About This User Guide

This user guide is a practical guide to using the *Relibase* and *Relibase+* tools for searching protein/ligand structures. It includes instructions on using the graphical user interface*, Hermes for Relibase* as well as providing help on relevant scientific issues.

Use the **<** and **>** navigational buttons above to move between pages of the user guide and the **TOC** and **Index** buttons to access the full table of contents and index. Additional on-line *Relibase+* resources can be accessed by clicking on the links on the right hand side of any page.

An extensive set of tutorials are also available for *Relibase+*. Tutorials can be accessed by clicking on the **Tutorials** link on the right hand side of any page.

The *Relibase+* user guide is divided into the following sections:

CHAPTER 1: THE RELIBASE+ DATABASE (see page 2)
CHAPTER 2: GENERAL FEATURES OF RELIBASE+ (see page 17)
CHAPTER 3: RUNNING RELIBASE+ SEARCHES (see page 57)
CHAPTER 4: RUNNING SIMILAR CAVITY SEARCHES (see page 97)
CHAPTER 5: USING THE RELIBASE SKETCHER (see page 119)
CHAPTER 6: CREATING IN-HOUSE DATABASES (see page 157)

CHAPTER 1: THE RELIBASE+ DATABASE

1 Coverage of the Relibase+ Database (see page 2)
2 Database Entries (see page 2)
3 Entry codes (see page 2)
4 Database Statistics (see page 3)
5 Database: Information content (see page 3)
6 A Typical Relibase+ Entry (see page 10)

1 Coverage of the Relibase+ Database

- The database behind Relibase+ is the Protein Data Bank (PDB) (*http://www.rcsb.org/pdb*). It covers all entries in the PDB which were determined experimentally by means of X-ray diffraction or NMR spectroscopy, but not theoretical structures. However, structures where a ligand (substrate) molecule was modelled into an experimental protein structure, *are* included.
- In Relibase+, all non-protein moieties in a structure are considered to be ligands. Hence metal ions, anions, solvate molecules (except water), cofactors and inhibitors are all regarded as ligands. In the 3D visualiser, DNA and RNA strands are displayed as ligands, but they are ignored in ligand-substructure searches.

2 Database Entries

Each protein entry in the Relibase+ database corresponds to an entry in the PDB and contains the following information (see Database: Information content Section 5, page 3):

- Bibliographic, textual and numerical information
- Crystal structure data (for X-ray structures)
- Protein chain(s)
- Binding site(s)
- Chemical diagram of the ligand(s)
- Crystal packing of the protein-ligand binding site
- Water structure information
- Cavity information
- Secondary structure

3 Entry codes

Relibase+ uses the same entry codes as the Protein Data Bank (PDB), i.e. one digit, followed by three characters (e.g. 1abe). All modifications to the entry codes, e.g. superseded entries, reflect those made to PDB.

4 Database Statistics

A list of summary statistics (number of PDB entries, number of ligand templates and number of ligand models) for the currently loaded databases can be found by following the *Database Statistics* link from the *Help* menu.

A list of all PDB entries which have been excluded from Relibase+, with associated reasons are available in:

```
$RELIBASE_ROOT/etc/Refused_PDB_entries.list
```

5 Database: Information content

The Relibase+ database contains all the information stored in the original PDB files. Searchable information fields are described in the following sections.

5.1 Bibliographic Information

- Authors' names
- Publication date
- Deposition date (Not searchable)

5.2 Textual Information

- The PDB HEADER, COMPND and SOURCE records
- Experimental method (X-ray or NMR)

5.3 Sequence Information

- Amino-acid sequence of protein chains

5.4 Chemical Information

- Ligand compound name
- Ligand entry code
- 2D chemical connectivity (used for 2D and 3D substructure searches, non-bonded interaction searches)
- Ligand Molecular weight

5.5 Crystallographic Information

- Unit cell parameters (Not searchable)
- Space group (Not searchable)

- Resolution
- Crystal packing of the protein-ligand binding site

5.6 Water Molecule Descriptors

Relibase+ includes a series of descriptors which are precalculated for each individual water molecule and stored in a database. The descriptors are described in the following sections.

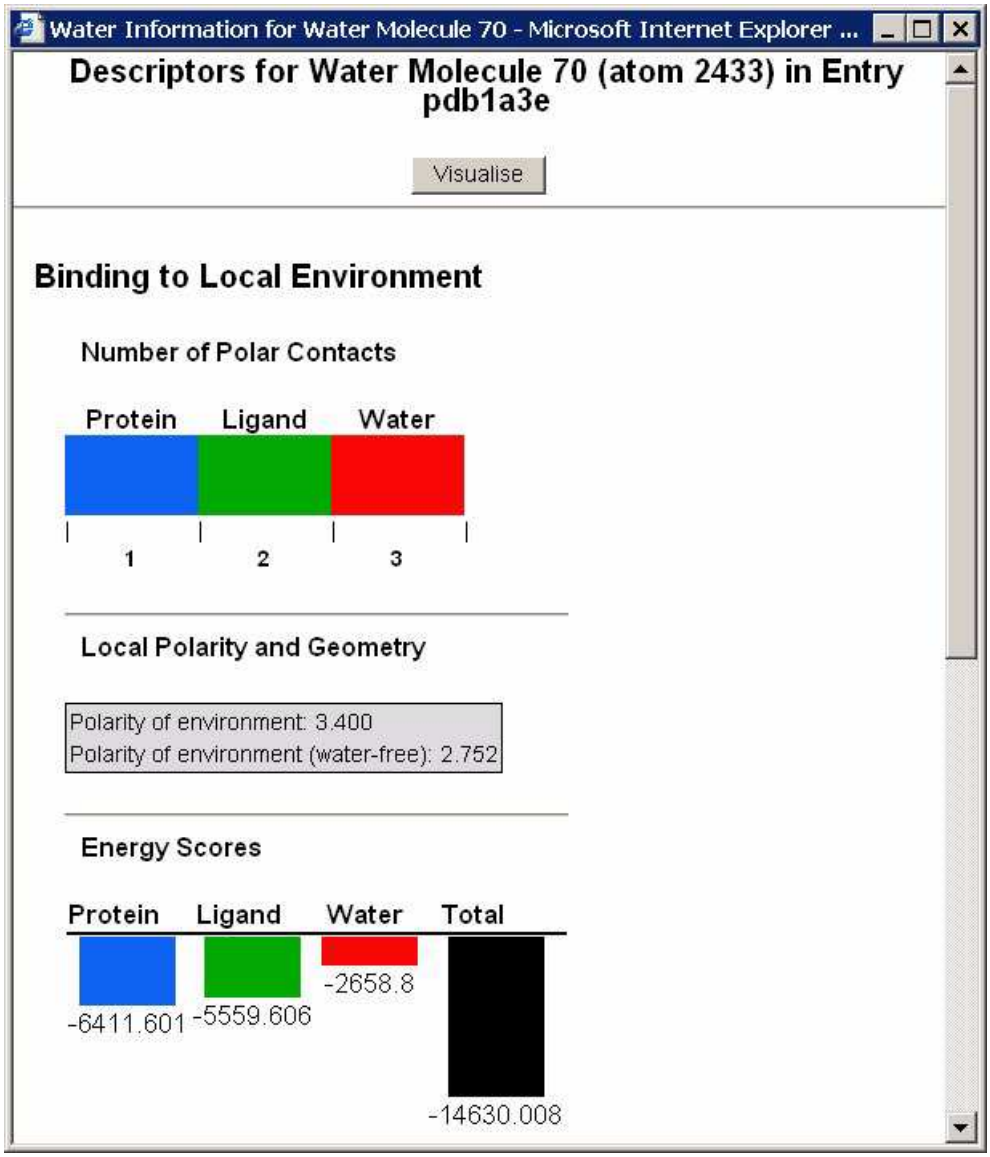
5.6.1 Binding of a Water to its Local Environment

The binding of water to its local environment can be described in terms of:

- Number of Polar Contacts (see page 4)
- Polarity of the Local Environment (see page 5)
- Coordination Geometry (see page 6)
- DrugScore Energy Score (see page 6)

Number of Polar Contacts

- This set of descriptors reports the number of polar atoms, i.e. the potential hydrogen bond partners within a 3.3Å radius of the water molecule in question. The total number is split up into polar contacts to protein, ligand and other water molecules respectively.
- Atoms taken into account are O, N, and Cl atoms, halide anions, metal cations. For O and N atoms, contact distances shorter than 2.4Å are considered but will also result in a warning (not for metal cations).
- The number of protein, ligand, and water contacts are displayed using a colour-coded bar:



Polarity of the Local Environment

- This is distance dependent and is a scaled measure of the polarity of the local environment within a 3.7Å radius of the water molecule in question. Two descriptors for the polarity are provided: water-containing and water-free:

  Water-containing: this is calculated from the water molecule of interest to protein atoms of type O, N, S, Cl atom, metal, halogen ion and aromatic carbon.

  Water-free: the same as water-containing, however protein water molecules within 3.7Å of the reference water molecule are not included in the summation.

- A linear cutoff function applies for specific atoms in the shell between 3.3Å and 3.7Å. The

atom-type weighting scheme is as follows:

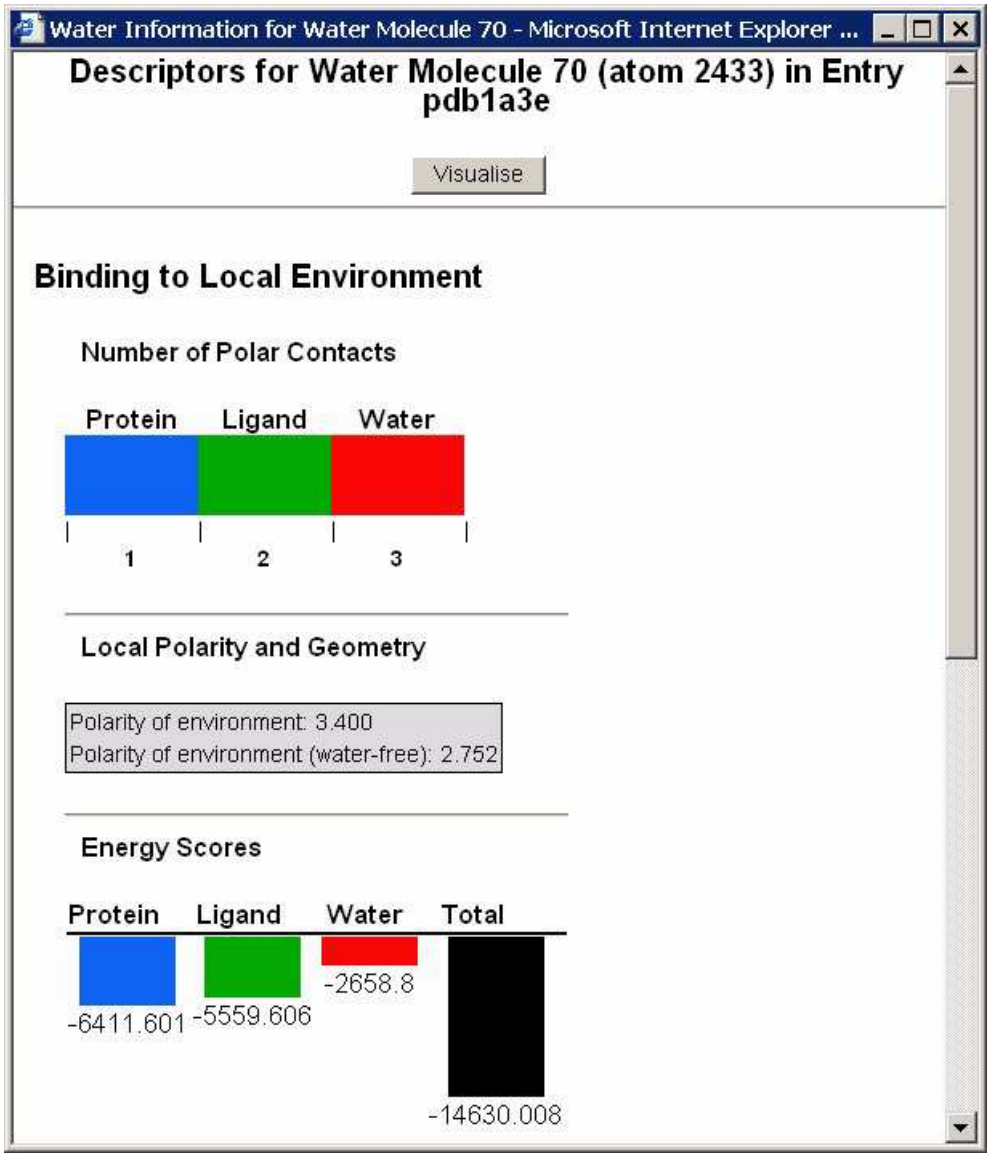
atomtype	w(atomtype)
O	1
N	1
S	0.5
C (arom)	0.15
metals	formal charge, but always <= 2
Cl	1 (always)
F,Br,I	1 (if anion)

Coordination Geometry

- If the water molecule has 4 or more polar atoms in its neighbourhood, i.e. potential hydrogen bond partners, the arrangement of these atoms is compared with an ideal tetrahedron; normalized bond lengths are used for all calculations. All permutations of 4-atom sets are superimposed onto an ideal tetrahedron, and the minimum RMS deviation is reported.
- The average deviation from tetrahedron angles in the observed polyhedron is also reported, however, no descriptor value is shown if the number of neighbouring atoms is 3 or less.

DrugScore Energy Score

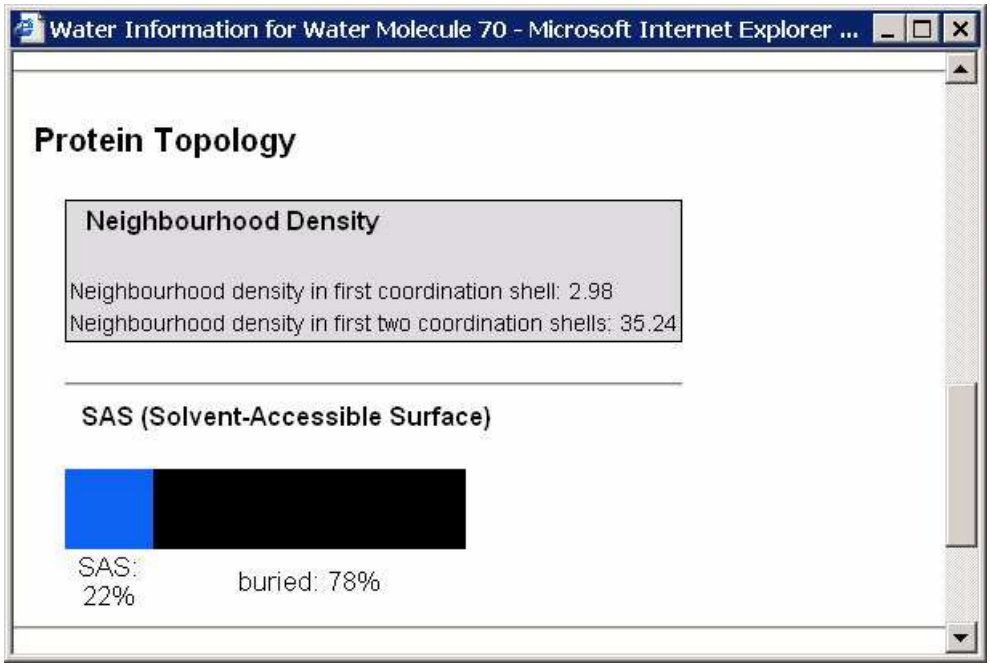
- The DrugScore energy score relates to the interaction energy between the water molecule and its local environment. The energy score is calculated from knowledge-based potentials which have been derived from the observed preferences for particular atom-pair interactions.
- By implementing a new atom type for water oxygen atoms, special potentials describing preferred types of water-protein and water-ligand interactions have been derived analogously to DrugScore (see References, page 172). All contact pairs up to a length of 6.0Å in a selected dataset were used to derive the potentials.
- The total score is calculated from the individual protein-water, ligand-water, and water-water contributions. All contributions are displayed in the form of coloured bars as shown below and the units are unscaled DrugScore units.



5.6.2 Local Topology of the Protein Structure

The protein topology is described in terms of:

- Neighbourhood Density (see page 8)
- SAS (Solvent-Accessible Surface) (see page 8)


Water Information for Water Molecule 70 - Microsoft Internet Explorer ...
Protein Topology
Neighbourhood Density
Neighbourhood density in first coordination shell: 2.98
Neighbourhood density in first two coordination shells: 35.24
SAS (Solvent-Accessible Surface)
SAS: 22%
buried: 78%

Neighbourhood Density

- This is a simple characterisation of the local degree of burial (micro-cavity) of a water molecule, based on analysing the local atom density.
- The descriptor reports the weighted sum of non-hydrogen atoms within the first (3.7Å) and second (7.0Å) coordination shells, respectively; water molecules are neglected in the summation. Linear cutoff functions apply between 3.3Å and 3.7Å, and between 6.5Å and 7.0Å, respectively.

SAS (Solvent-Accessible Surface)

- This descriptor represents the portion of the water sphere (VdW radius 1.4Å) which is not covered by protein or ligand atoms. It is a more refined characterisation of the degree of burial, compared to the neighbourhood density. For the calculation the water molecule is treated as a ligand or protein oxygen atom, with all other water molecules excluded.

5.6.3 Data-Related Issues

The following data are available for individual water molecules:

- Crystallographic B-factor.
- Mean B-factor of protein environment. All protein atoms within a 3.3Å radius are considered.
- Mean B-factor of environment.
- Crystallographic occupancy.
- Mobility, a scaled measure for the mobility of a water molecule encountering the

crystallographic occupancy of the water molecule, and the average level of B-factors and occupancies in the structure (see References, page 172).

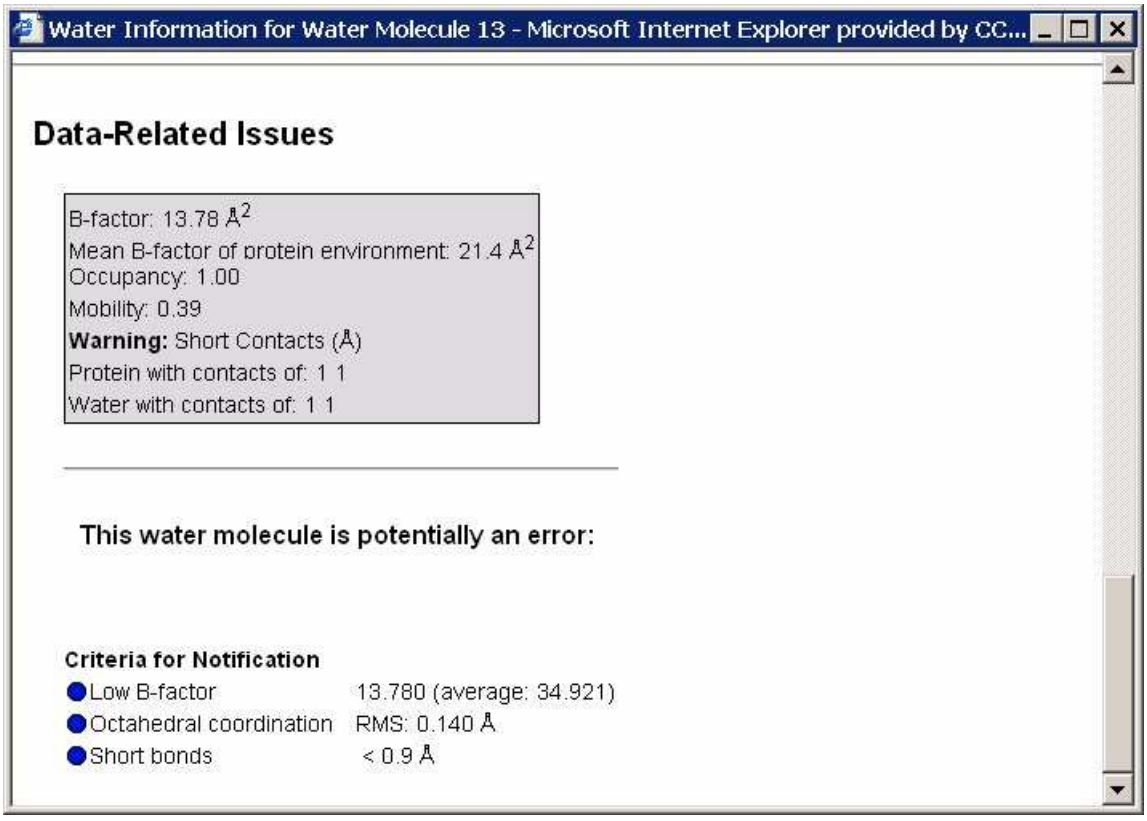
Mobility(i) = (B-fac(i) / <B-fac>) / (occ(i) / <occ>)

- Short contacts. All contacts shorter than 2.4Å are reported as warnings. This doesn't mean there is an error in the structure, it only points the user to a potential problem.
- An almost octahedral coordination of a water molecule indicates a crystallographically misassigned atom which is more likely to be a Na or Mg atom.

The criteria used for notification of a potentially erroneous water molecule are:

- The B-factor of the water molecule is below 20Å^2, or below the average B-factor in the structure.
- There are short contacts to O and/or N atoms, leading to a high valence (see References, page 172).
- The RMS deviation between the best fitting polyhedron and an ideal octahedron is < 0.25Å.

A combination of these criteria are used to decide upon whether a water molecule is notified as being dubious. The criteria for notification are displayed as shown below:


Water Information for Water Molecule 13 - Microsoft Internet Explorer provided by CC...
Data-Related Issues
B-factor: 13.78 Å2
Mean B-factor of protein environment: 21.4 Å2
Occupancy: 1.00
Mobility: 0.39
Warning: Short Contacts (Å)
Protein with contacts of: 1 1
Water with contacts of: 1 1
This water molecule is potentially an error:
Criteria for Notification
Low B-factor 13.780 (average: 34.921)
Octahedral coordination RMS: 0.140 Å
Short bonds < 0.9 Å

5.7 Cavity Information

- All cavities in a protein structure are listed, with their volumes and any ligands they contain.

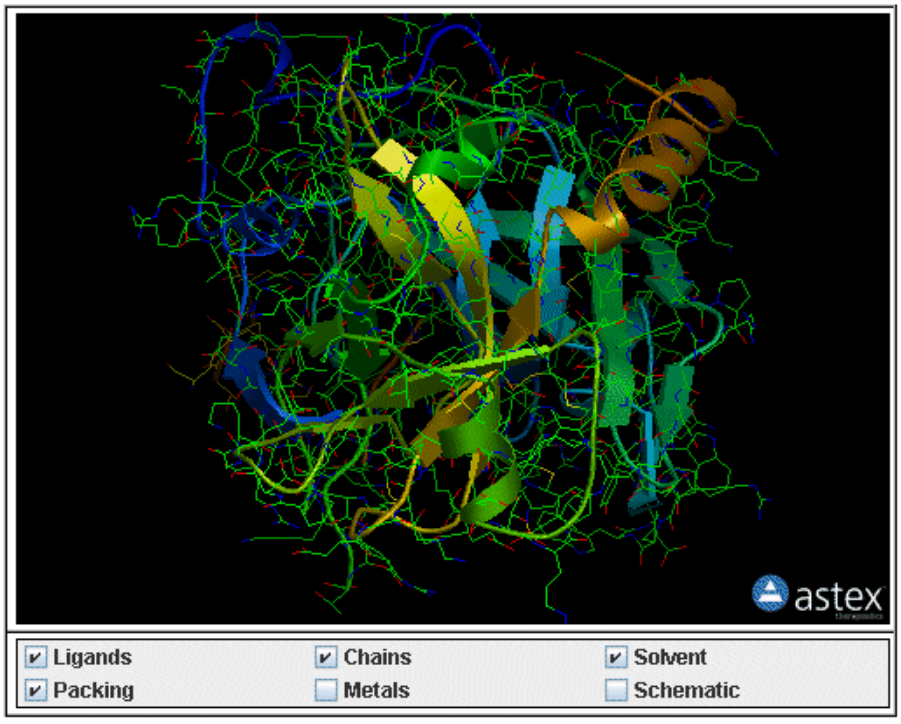
  Note: Very large cavities (> 3000Å^3) are usually of little interest (they are often ill-defined gaps between large protein domains).
- Cavity information for any database entry can be accessed by clicking on the **Cavity Information** button at the bottom of either Protein or Ligand Information pages (see Accessing Cavity Information for Relibase+ Database Entries Section 2, page 99).

5.8 Secondary Structure Information

- Details of helices, beta-sheets and turns in the protein can be viewed and displayed in 3D via the Secondary Structure Information button at the bottom of each Protein Information page. (see Secondary Structure Information Section 4.6, page 35).

6 A Typical Relibase+ Entry

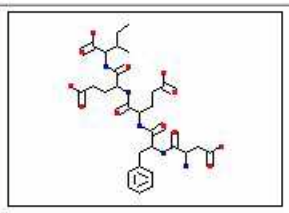
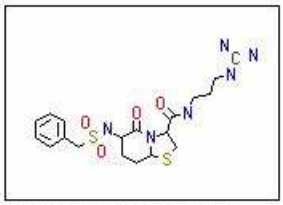
Embedded 3D visualisation via AstexViewer:



Bibliographic and chemical text information:

Header	SERINE PROTEASE
Title	X-RAY STRUCTURE OF THE COMPLEX OF HUMAN ALPHA THROMBIN WITH THE INHIBITOR SDZ 229-357
Compound	MOL_ID: 1 MOLECULE: ALPHA THROMBIN CHAIN: A, B, F, E EC: 3.4.21.5
Reference	J.MED.CHEM. **41**, (1998), 3664
Author(s)	J.KALLEN
Source	MOL_ID: 1 ORGANISM_SCIENTIFIC: HOMO SAPIENS ORGANISM_COMMON: HUMAN
Method	X-RAY DIFFRACTION
Crystal	Cell: a = 71.000, b = 72.200, c = 73.100, alpha = 90.000, beta = 100.700, gamma = 90.000, Spacegroup = C121
Resolution	2.300
RFactor	Unavailable
Deposition Date	10.6.1998

2D diagrams of ligands:

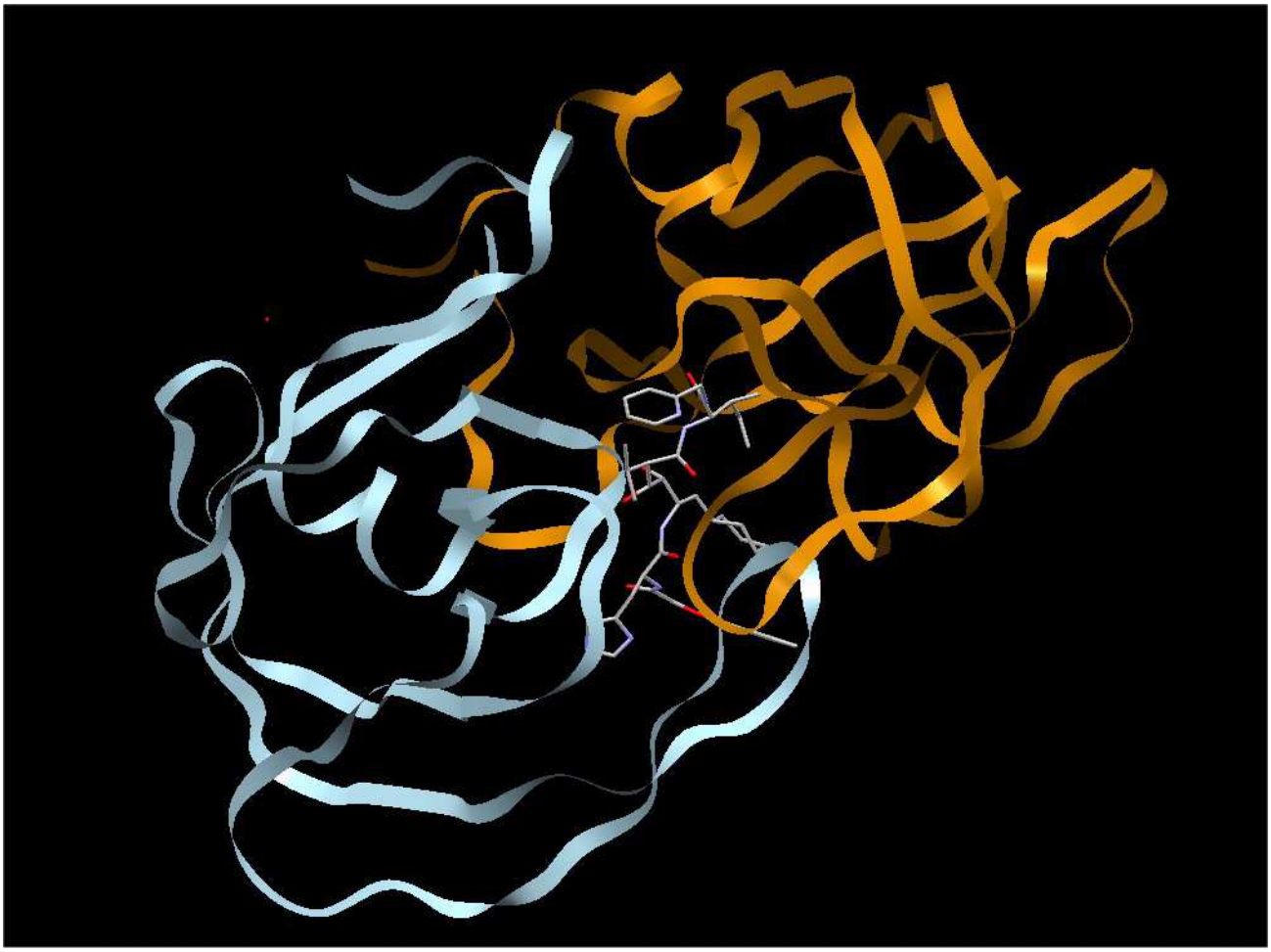
Ligands	 DFE... in PDB-Entry 1bhx **Water-Mediated Protein-Ligand Contact Paths: 3**  R56 in PDB-Entry 1bhx **Water-Mediated Protein-Ligand Contact Paths: 3**

Sequence data (which can be used in a Similar Chain Search):

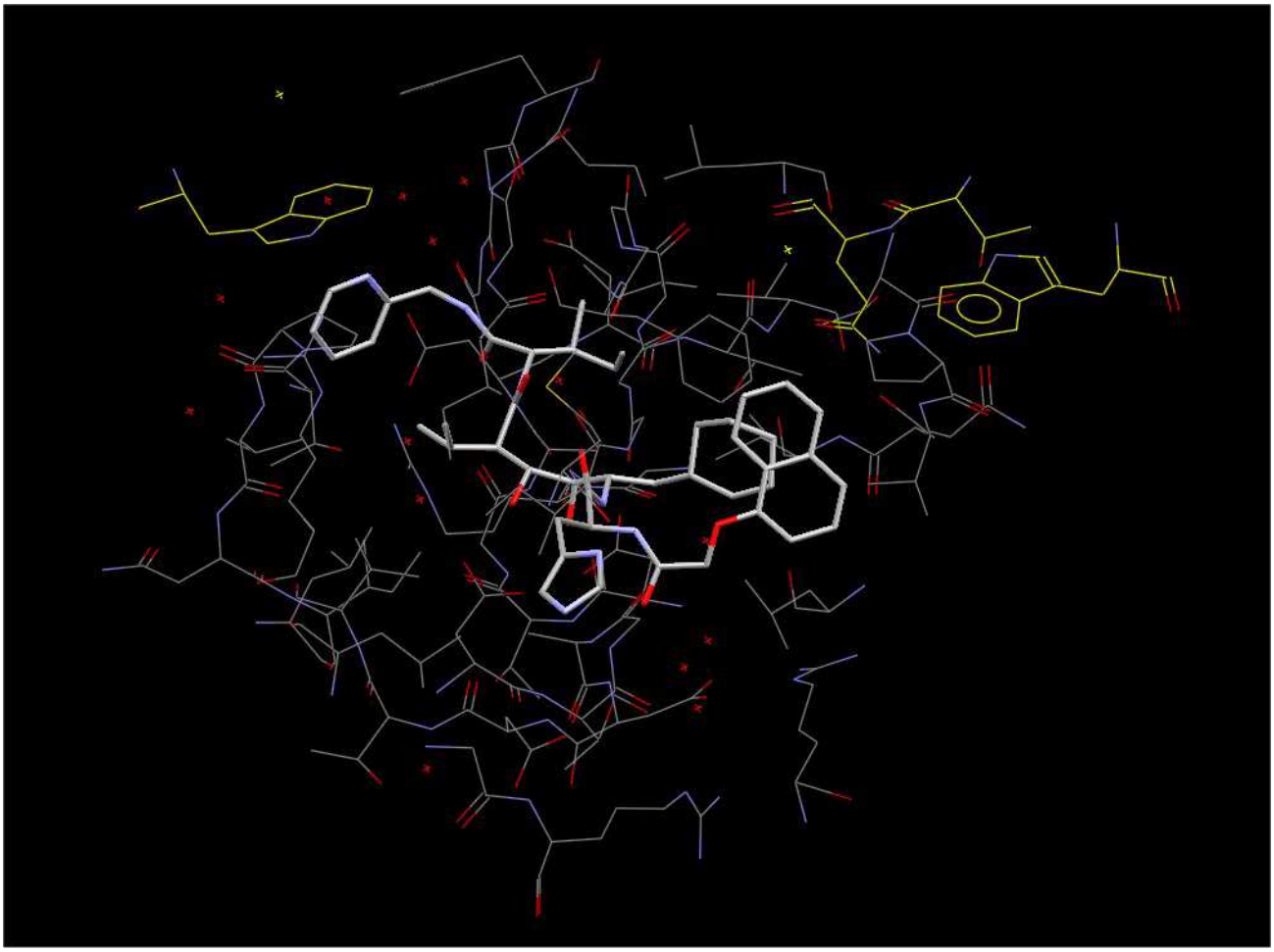
Protein Chain pdb1bhx-A

SGEADCGLRPLFEKKSLEDKTERELLESYI

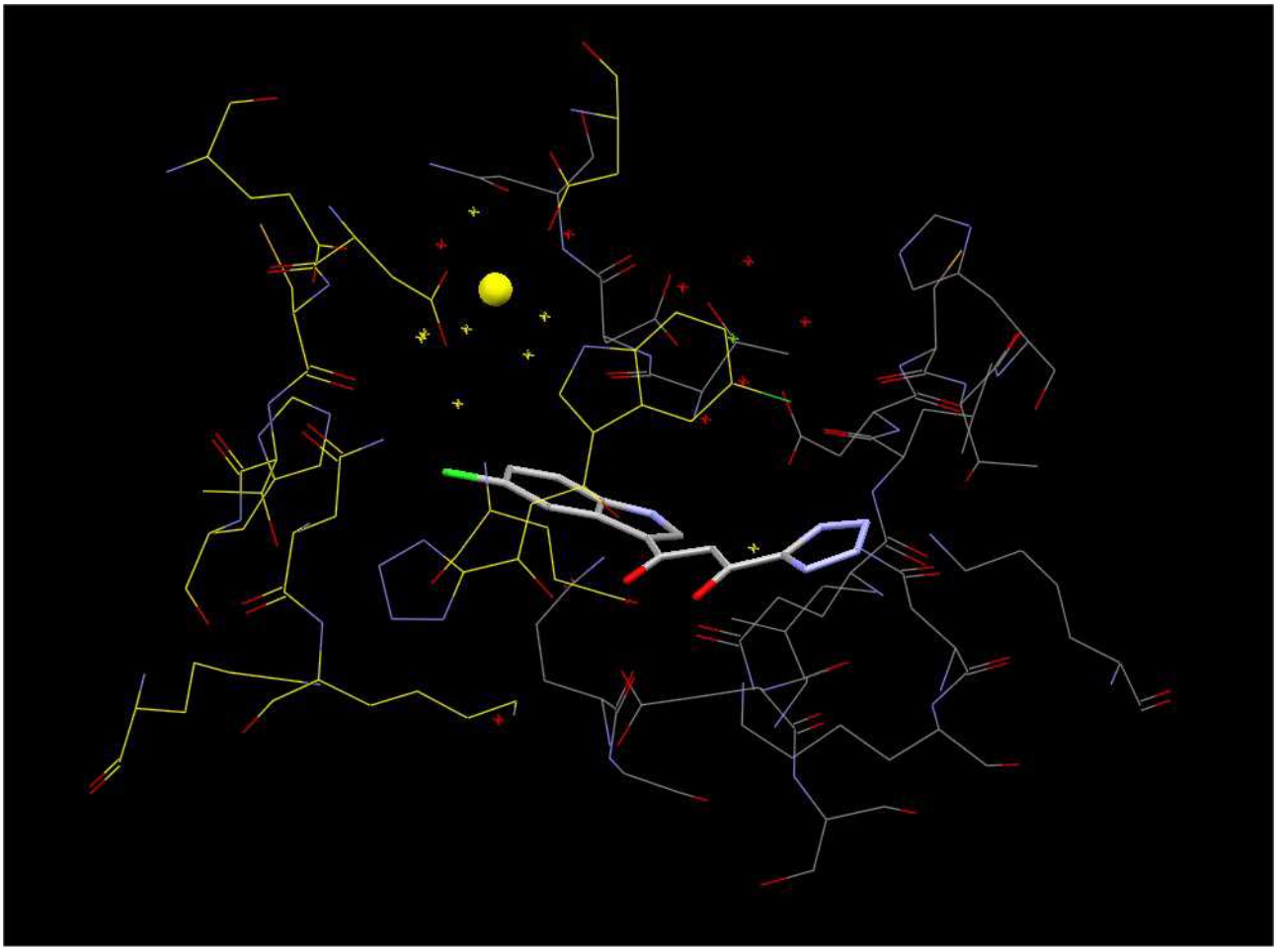
Comprehensive 3D visualisation and exploration via Hermes (the protein structure of 1hiv is shown below):



3D structure of binding site (1hiv):



Crystal packing of protein-ligand binding site (1qs4):



Information on water structure and water-mediated protein-ligand contacts:

General Information on Water Structure in Entry pdb1a3e

Water models:	1
Water molecules:	206
Water molecules per residue:	0.741
Average B-factor:	34.921 Å^2
Standard deviation of B-factors:	10.231

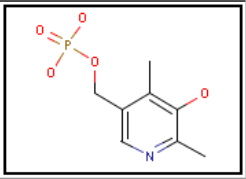
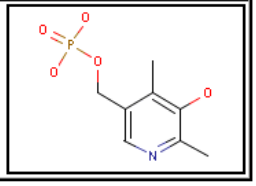
Dubious Water Molecule

Criteria for Notification 13

Low B-factor	13.780 (average: 34.921)
Octahedral coordination	RMS: 0.140 Å
Short bonds	< 0.9 Å

Cavity information:

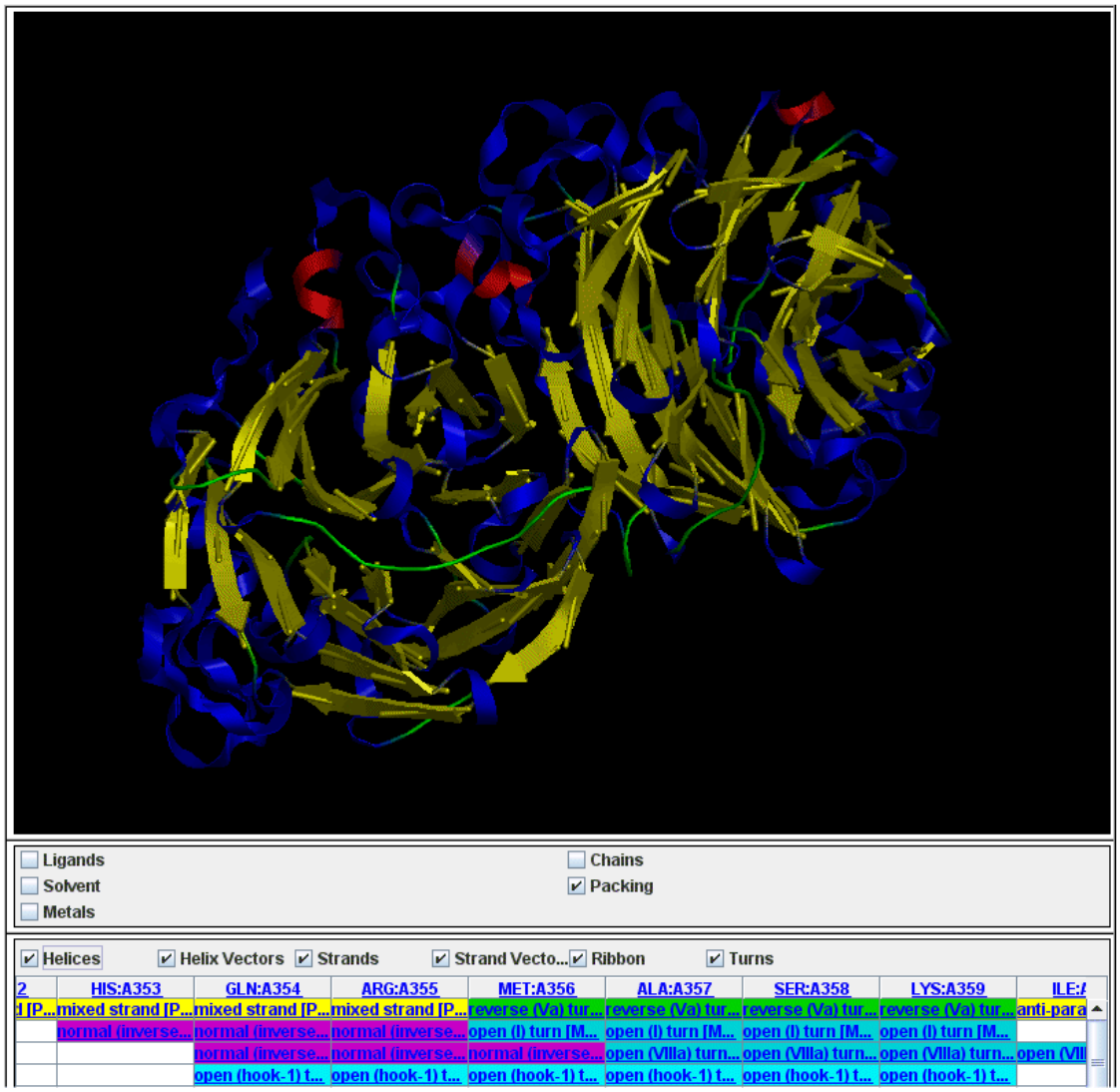
Cavities in PDB Entry 8aat

Cavity	Volume ($Å^3$)	Ligand
pdb8aat.1	310.5	
pdb8aat.2	402.6	
pdb8aat.3	422.0	
pdb8aat.4	489.9	
pdb8aat.5	587.2	
pdb8aat.6	738.4	
pdb8aat.7	999.4	
Warning: Cavities larger than 3000Å many be slow to load in the visualiser and are unsuitable for cavity searching.		
pdb8aat.8	3387.2	
pdb8aat.9	3541.2	

Similarity Search Results

Information on secondary structure:

Secondary Structure Information for PDB Entry 1a4g



CHAPTER 2: GENERAL FEATURES OF RELIBASE+

1 Getting Started (see page 17)
2 Starting a Search (see page 19)
3 Viewing and Navigating Search Results (see page 20)
4 Viewing Information for Individual Hit Structures (see page 27)
5 3D Visualisation of Structures (see page 42)
6 Storing, Combining and Converting Search Results (see page 47)

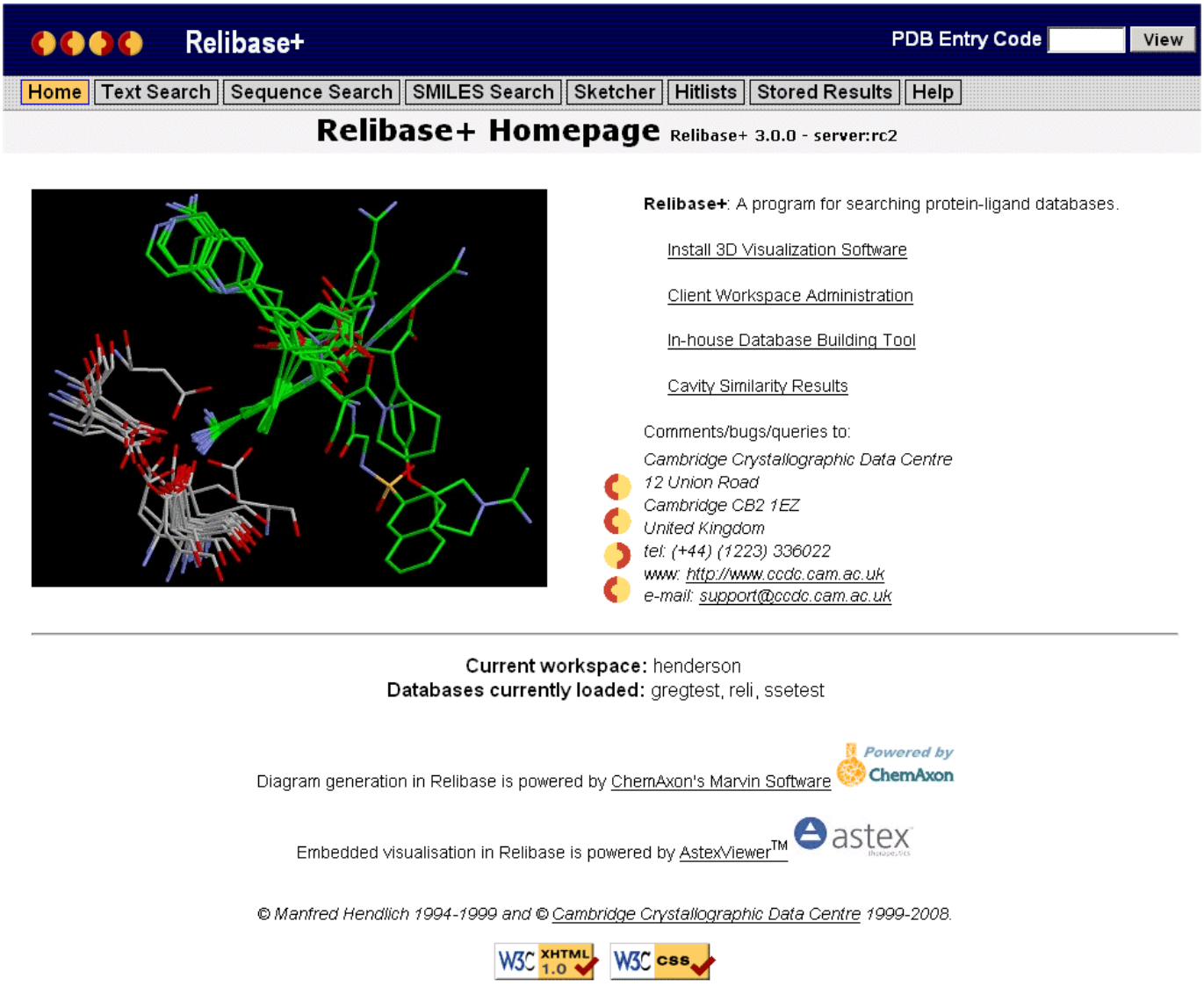
1 Getting Started

1.1 The Basics of Using Relibase+

- Relibase+ is a web-based application and all its functionality is accessible via a web browser. The main Relibase+ page features a menubar which contains the following buttons:
  - **Home**: provides access to the Relibase+ Home Page (see The Relibase+ Home Page Section 1.2, page 17).
  - **Text Search, Sequence Search, SMILES Search, Sketcher:** provide access to query constructor pages (used to define queries and start searching the Relibase+ database).
  - **Hitlists:** used to view and combine results from previous Relibase+ searches (see Storing, Combining and Converting Search Results Section 6, page 47).
  - **Stored Results**: used to access results from previous searches, binding site superpositions (see Similar Binding Site Searches (and Superposition) Section 9, page 84), similar sequence searches (see Protein Sequence Searches Section 4, page 65) and cavity similarity searches (see Cavity Similarity Searching Section 4, page 107).
  - **Help:** provides access to the Relibase+ User Guide and technical documentation.

1.2 The Relibase+ Home Page

- When you first access the Relibase+ server, the Relibase+ Home Page will be displayed. The Home Page can also be accessed from any Relibase+ page by hitting the **Home** button in the top menubar.



- The Home Page provides the following options:
  - Click on the CCDC logo to go to the CCDC web site (*http://www.ccdc.cam.ac.uk*).
  - Enter a PDB code into the *PDB Entry Code* box and click **View** for quick access to a protein of interest.
  - Click on the link **Install 3D Visualization Software** to download *Hermes*, the software required to visualise Relibase+ entries in 3D.
  - Click on the **Client Workspace Administration** to access other client workspaces (available for unlimited licenses only). The workspace username and the databases that are currently loaded are displayed at the bottom of the page.
  - Click on the link **In-house Database Building Tool** to build proprietary database(s).
  - Use the **Cavity Similarity Results** link to hyperlink to previously saved cavity hitlists.
  - Click on *support@ccdc.cam.ac.uk* to email us with any problems, enhancement requests etc.

2 Starting a Search

The buttons on the Relibase+ menubar provide access to query constructor pages. In these pages you can define queries and start searches. Note that the *PDB Entry Code* box can be used for quick access to a specific PDB code of interest.

- **Text Search** provides access to**:**
  - Searches on protein entry code (see PDB Entry Code Searches Section 2, page 57).
  - Text searches (see Keyword Searches Section 3, page 58).
  - Author name searches (see Author Name Searches Section 3.2, page 59).
  - Ligand compound name searches (see Ligand Compound Name Searches Section 3.3, page 61).
  - Ligand entry code searches (see Ligand Code Searches Section 3.4, page 62).
  - Database browsing capabilities (see Browsing Database Entries Section 1, page 57).
- **Sequence Search** provides access to**:**
  - Searches on amino acid sequences (see Protein Sequence Searches Section 4, page 65).
- **SMILES Search provides access to:**
  - Searches on ligand SMILES or SMARTS strings (see Ligand SMILES or SMARTS Searches Section 5, page 66).
- **Sketcher** provides access to**:**
  - 2D/3D ligand substructure searching (see 2D/3D Ligand Substructure Searches Section 6, page 70).
  - Non-bonded (protein-ligand or protein-protein) interaction searching (see Non-bonded interaction searching Section 6.6, page 72).

Some Relibase+ searches can only be started from Relibase+ protein entry pages (see Relibase+ Protein Entry Pages Section 4.1, page 27), or from Relibase+ ligand pages (see Relibase+ Ligand Pages Section 4.2, page 28), these include:

- Ligand similarity searching (see Similar Ligand Searches Section 7, page 78).
- Similar chain searching (see Similar Protein Chain Searches Section 8, page 81).
- Similar binding site searching (and superposition) (see Similar Binding Site Searches (and Superposition) Section 9, page 84).

3 Viewing and Navigating Search Results

3.1 Overview

Relibase+ searches started from the Relibase+ menubar, i.e. **Text Search**, **Sequence Search**, **SMILES Search** and **Sketcher** searches (see Starting a Search Section 2, page 19), can result in a three different browsable lists of hits:

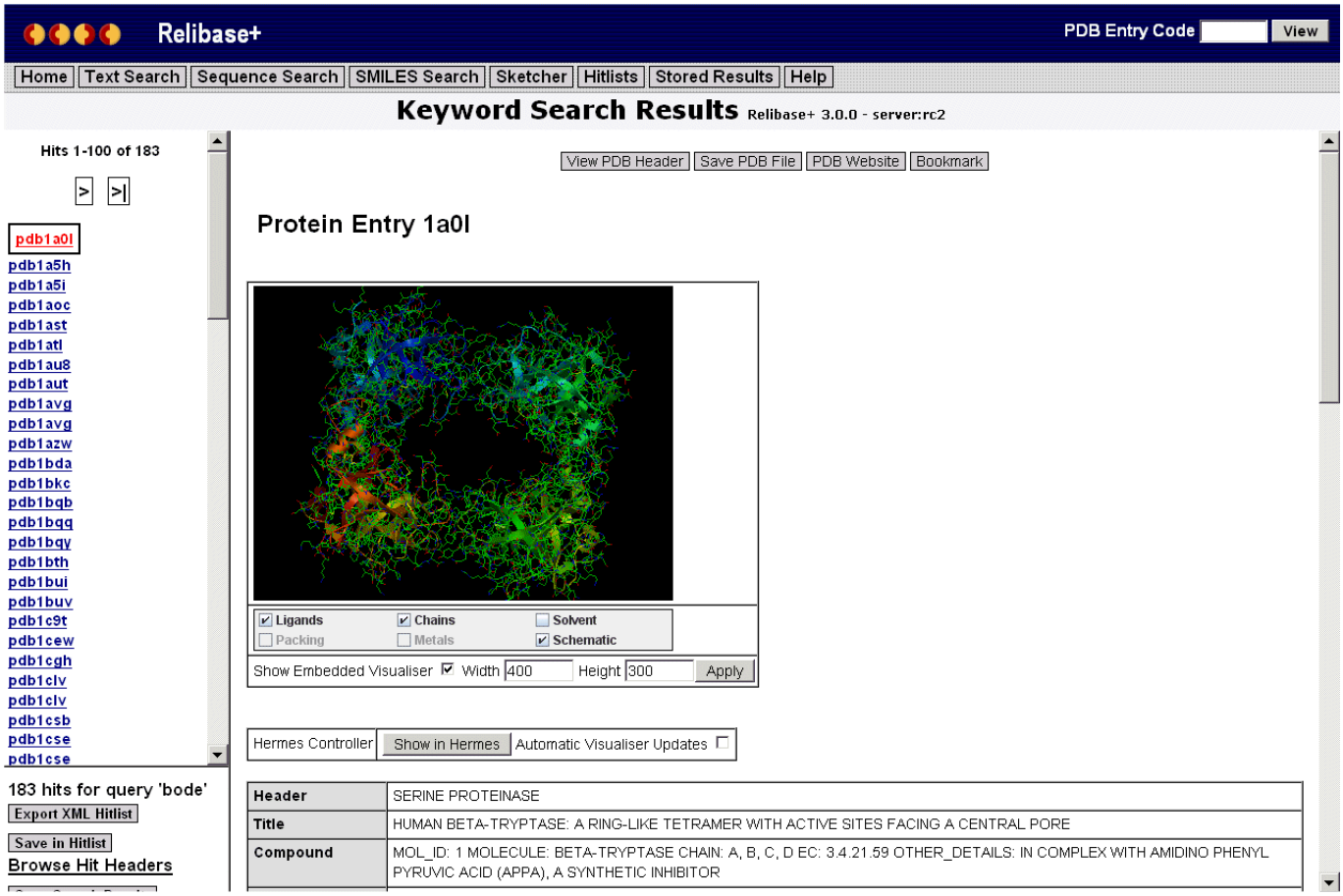
- A list of Relibase+ entries (see Using the Protein Entry Browser Section 3.2, page 20).
- A list of protein chains (see Viewing Sequence-Based Search Results Section 3.4, page 24)
- A list of ligands, with their binding sites (see Relibase+ Ligand Pages Section 4.2, page 28).

To view the results of other types of searches, started from Relibase+ protein entry pages (see Relibase+ Protein Entry Pages Section 4.1, page 27) or from Relibase+ ligand pages (see Relibase+ Ligand Pages Section 4.2, page 28), the user is referred to the sections where these searches are described:

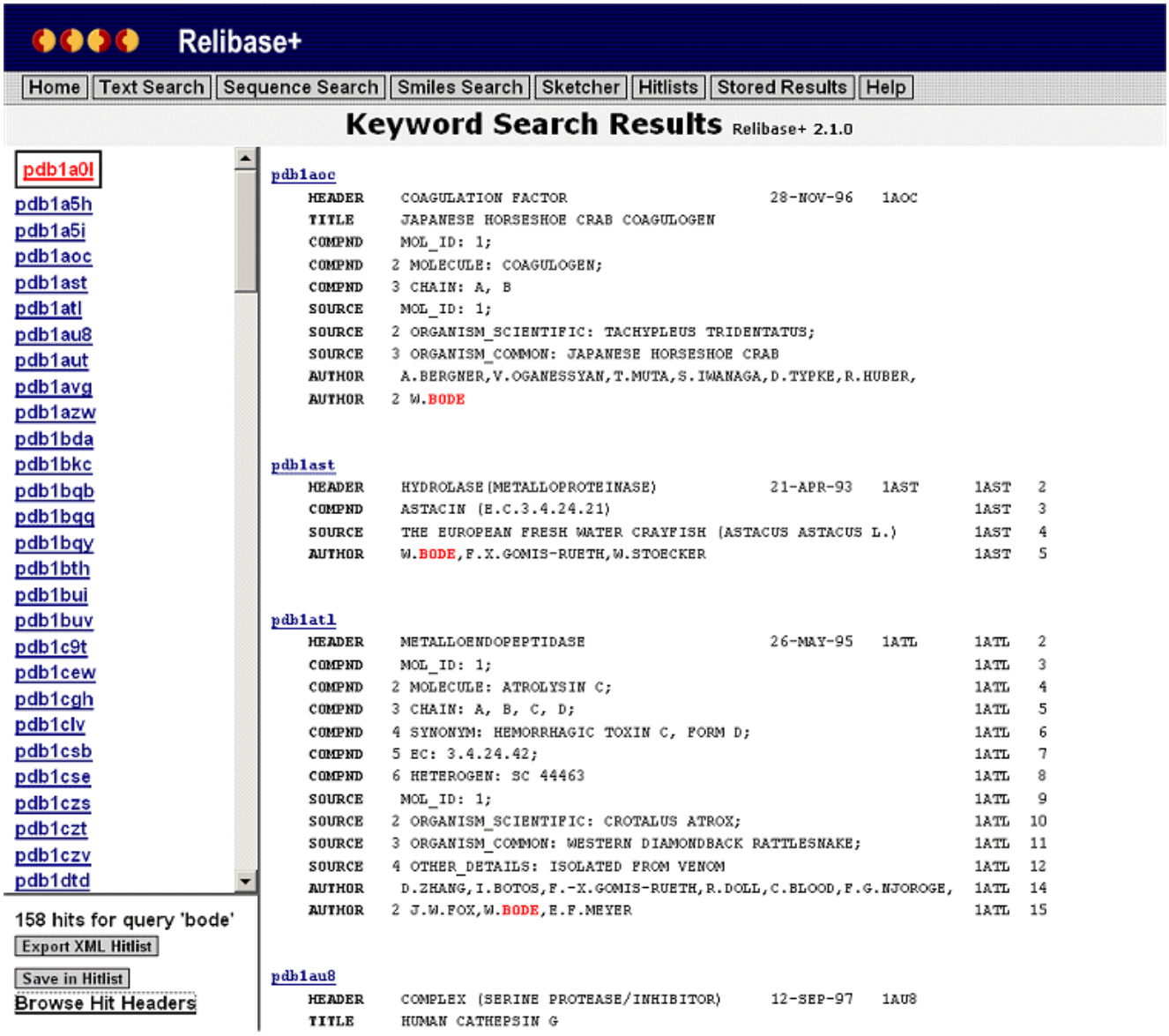
- Ligand similarity searching (see Similar Ligand Searches Section 7, page 78).
- Similar chain searching (see Similar Protein Chain Searches Section 8, page 81).
- Similar binding site searching (and superposition) (see Similar Binding Site Searches (and Superposition) Section 9, page 84).

3.2 Using the Protein Entry Browser

- Search results from text searches (see Keyword Searches Section 3, page 58) and author searches (see Author Name Searches Section 3.2, page 59) are displayed as three frames. The top-left frame contains the protein entry codes of the hits. The right-hand frame contains the Relibase+ protein entry page (see Relibase+ Protein Entry Pages Section 4.1, page 27) for the currently selected entry from the list (by default, this will be the first hit). Entries can be selected and inspected by clicking on the protein entry codes in the top-left frame.
- The following example shows the results for an author search for *bode*:



- The bottom-left frame displays the total number of hits found for the search. To get a better impression of the type of hits that were found, click on **Browse Hit Headers**. This will list the headers of the entries that were hit.
- The places where the query string was found are highlighted in red. Each protein entry code is linked to the corresponding Relibase+ entry page (see Relibase+ Protein Entry Pages Section 4.1, page 27).



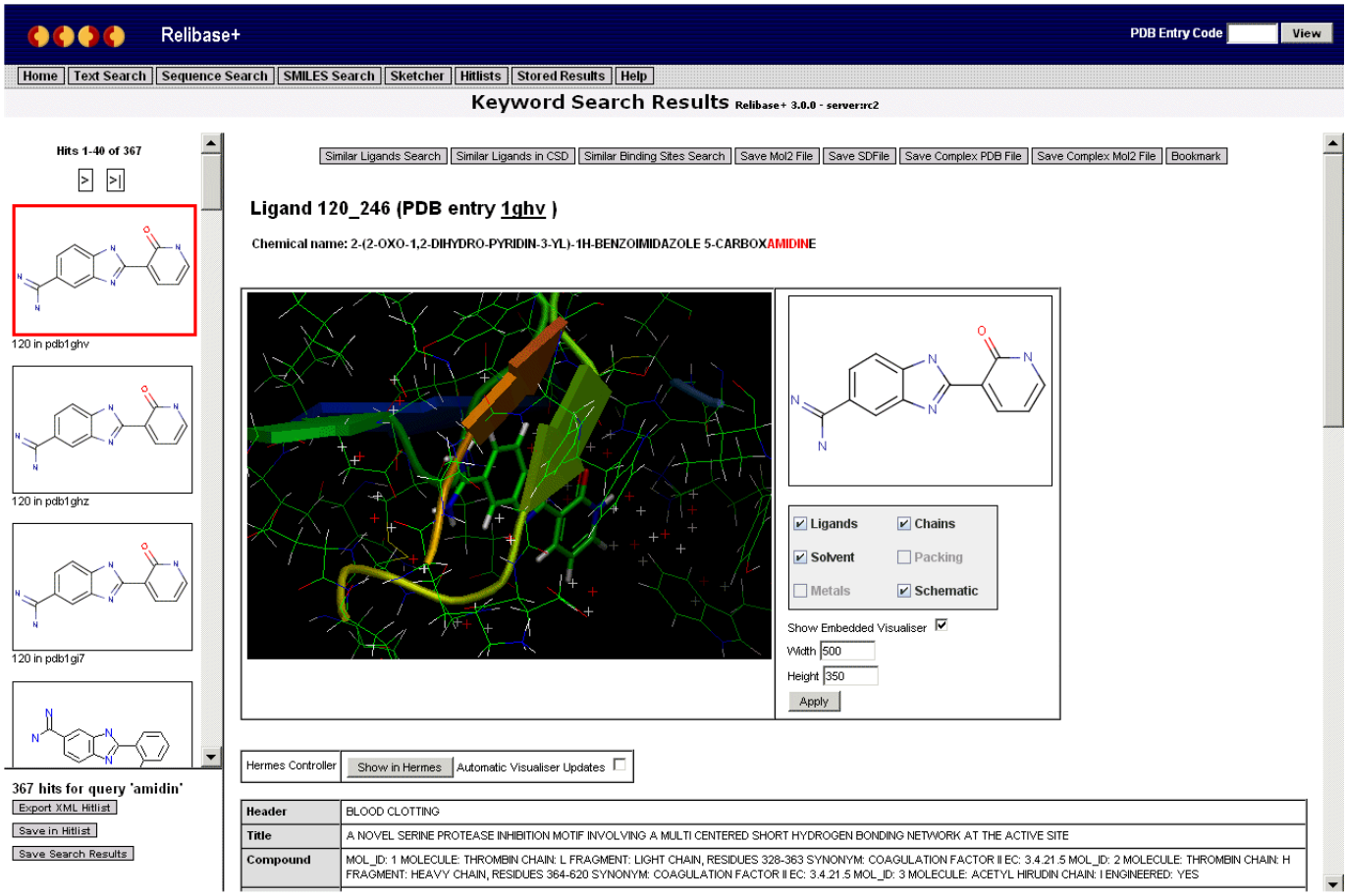
- To save the hitlist of protein entry codes as an *XML* format file, click on **Export XML Hitlist**. In the resulting pop-up window enter a name for the exported hitlist, then click **OK**.
  Note: any saved XML hitlist can be read back into Relibase+ (see Loading XML Format Hitlists Section 6.8, page 55).
- To save the hitlist of protein entry codes on the Relibase+ server, click on the **Save in Hitlist** hyperlink. In the resulting pop-up window, enter a name for the hitlist then click **OK**.

3.3 Using the Ligand Browser

- Search results for ligand-based searches, e.g. ligand name, ligand entry code, Smiles and Sketcher searches (see Starting a Search Section 2, page 19), are displayed as three frames. The top-left frame contains the 2D chemical diagrams of the ligands that matched the query. The right-hand frame contains the Relibase+ ligand page (see Relibase+ Ligand Pages Section 4.2,

page 28) for the ligand currently selected from the list (by default, the right-hand frame contains the ligand page of the first hit). Ligands can be selected and inspected by clicking on the diagrams in the top-left frame.

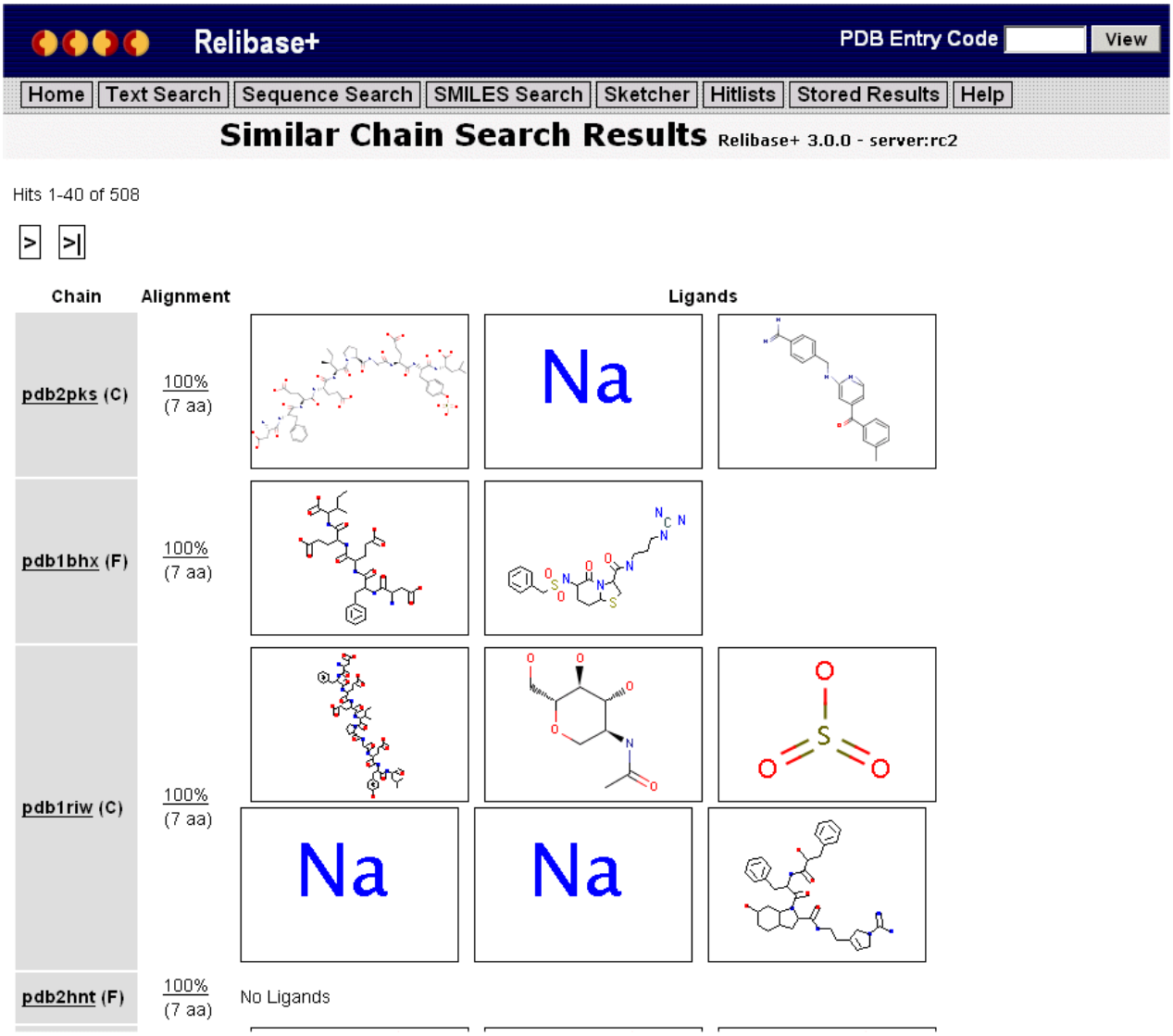
- The following example shows the results of a ligand name search for *amidin*:



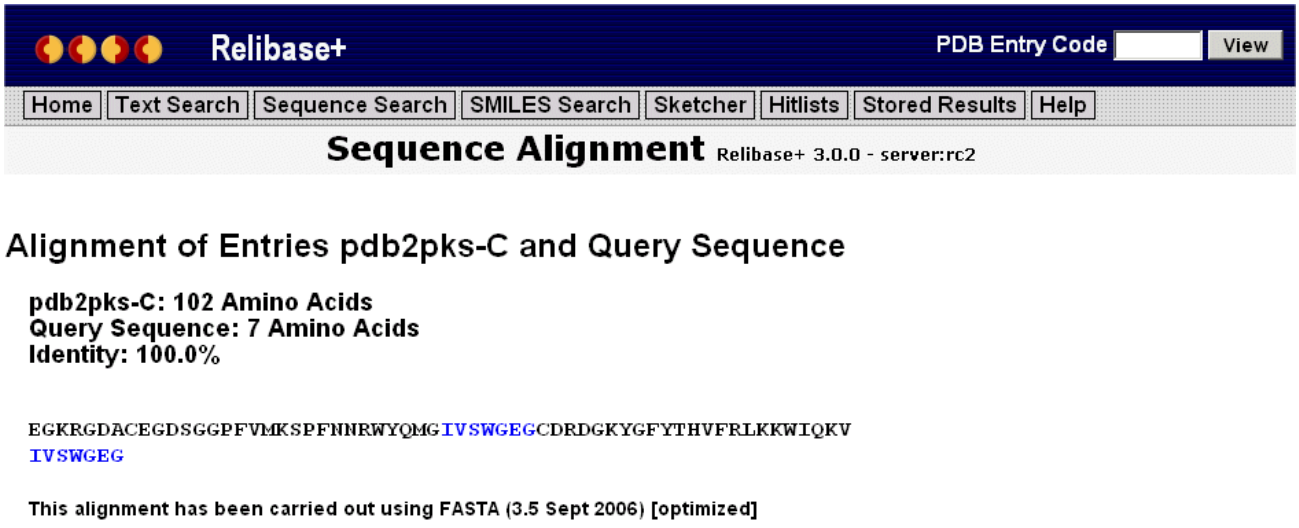
- The bottom-left frame displays the total number of hits found for the search. To save the hitlist of ligands as an *XML* format file, click on **Export XML Hitlist**. In the resulting pop-up window enter a name for the exported hitlist, then click **OK**.
  Note: any saved XML hitlist can be read back into Relibase+ (see Loading XML Format Hitlists Section 6.8, page 55).
- To save the hitlist of ligands on the Relibase+ server, click on the **Save in Hitlist** hyperlink. In the resulting pop-up window, enter a name for the hitlist then click **OK**. All ligands in a ligand hitlist can be saved in a multi-mol2 or SD file via the hitlists page (see Saving Hitlists Section 6.6, page 54).
- The search can be saved via the **Save Search Results** button. The search can be reloaded at a later date via the *Stored Results* window (see Storing Search Results Section 6.1, page 47).

3.4 Viewing Sequence-Based Search Results

- The results of sequence-based searches are displayed as a table of chains:



- The first column lists the Relibase+ protein entry code, together with a chain identifier, e.g. entry *pdb2pks* chain *C*. Click on the entry code to go to the Relibase+ protein entry page (see Relibase+ Protein Entry Pages Section 4.1, page 27) for this structure.
- The second column lists the percentage identity with respect to the reference chain and the length of the matched sequence. Above, in *pdb2pks*, a sequence match of seven amino acids (*7 AA*) has been made with the query sequence. For sequence searches (see Protein Sequence Searches Section 4, page 65) the reference chain is that part of the sequence typed into the sequence search box.
- Click on the percentage identity to analyse the alignment. For a typical sequence search the results look like this:



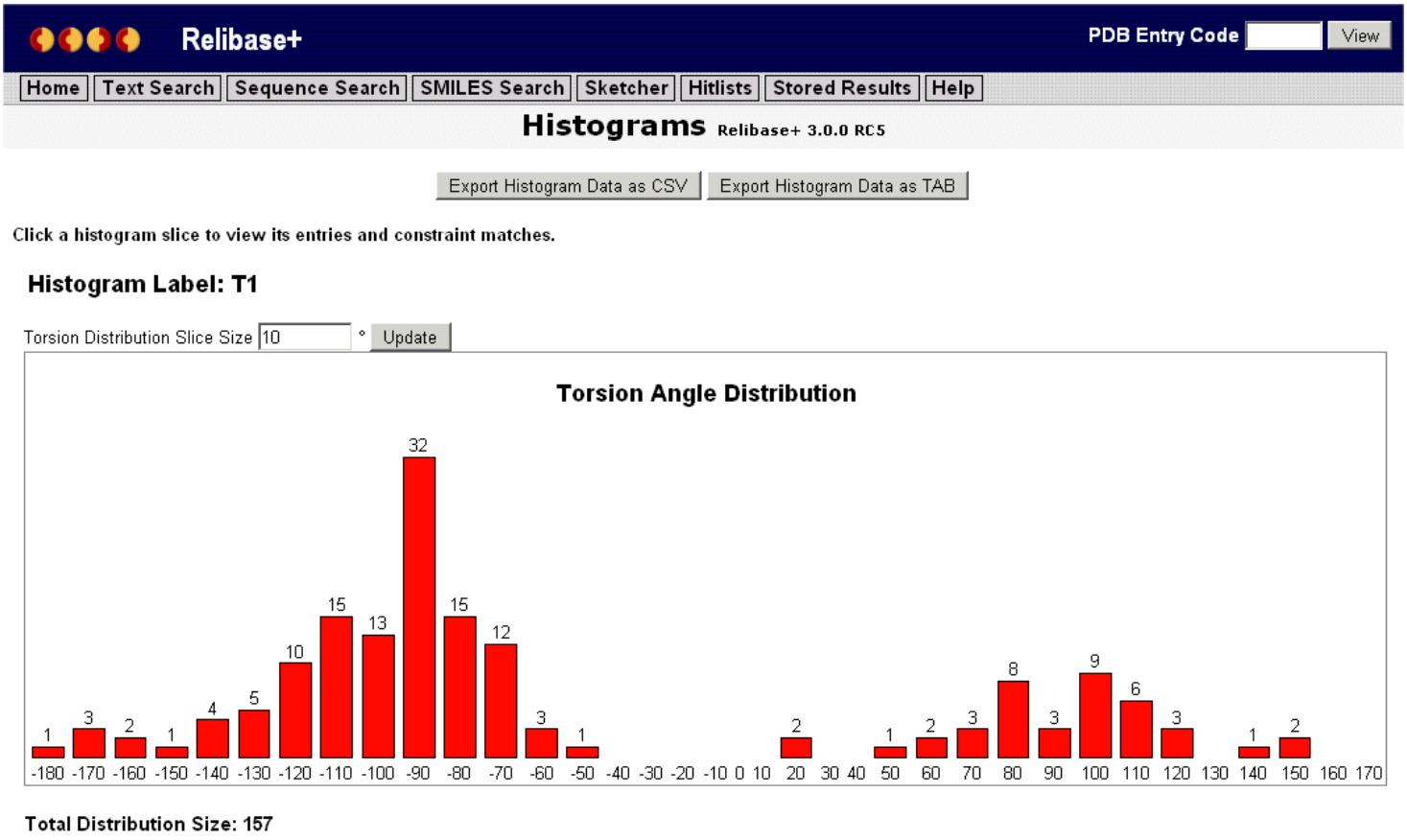
- The percentage *Identity* shown on the *Alignment of Entries* page (100.0% above) is for the entire hit protein compared to the query protein, determined by ALIGN.
- The ligand diagrams in the listing of chains are only shown if **Show Ligands** was selected when the search was started (see Protein Sequence Searches Section 4, page 65). Click on the diagrams to launch the individual ligand pages in a separate browser window (see Relibase+ Ligand Pages Section 4.2, page 28).

3.5 Viewing 3D Substructure Search Results; Geometrical Analysis

3.5.1 Viewing Distribution Histograms for Geometrical Parameters

- When you have defined geometrical parameters in your query (see Geometric Parameters Section 11, page 139), histograms for these parameters are generated automatically.
- The histogram(s) can be viewed by clicking on the **Histogram(s)** link in the bottom left frame of the Relibase+ ligand browser (see Using the Ligand Browser Section 3.3, page 22).

- This loads the histogram(s) into the browser:


Relibase+
PDB Entry Code
View
Home
Text Search
Sequence Search
SMILES Search
Sketcher
Hitlists
Stored Results
Help
Histograms
Relibase+ 3.0.0 RC5
Export Histogram Data as CSV
Export Histogram Data as TAB
Click a histogram slice to view its entries and constraint matches.
Histogram Label: T1
Torsion Distribution Slice Size 10 ° Update
Torsion Angle Distribution
1 3 2 1 4 5 10 15 13 32 15 12 3 1 2 1 2 3 8 3 9 6 3 1 2
-180 -170 -160 -150 -140 -130 -120 -110 -100 -90 -80 -70 -60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
Total Distribution Size: 157

- By default, histograms for angles and torsion angles are binned at 10 degree intervals, those for distances at 0.2Å intervals. To alter the distribution bin size enter a *distribution slice size* and hit **Update**. The histogram will be updated to reflect the specified bin size.
- The number of observations in each interval is shown at the top of each bar. Click on the individual bars to load all hits that make up that bar into the Relibase+ ligand browser (see Using the Ligand Browser Section 3.3, page 22).
- Specified parameter values for hit structures can be saved to a file for later analysis. Choose from:
  - **Export Histogram Data as CSV:** outputs the current histogram in .CSV file format.
  - **Export Histogram Data as TAB**: outputs the current histogram in .tab format, suitable for input to Vista (the statistical analysis package distributed with the Cambridge Structural Database System: *http://www.ccdc.cam.ac.uk/products/csd_system/vista/*.

3.5.2 Viewing 3D Superposition of Hits

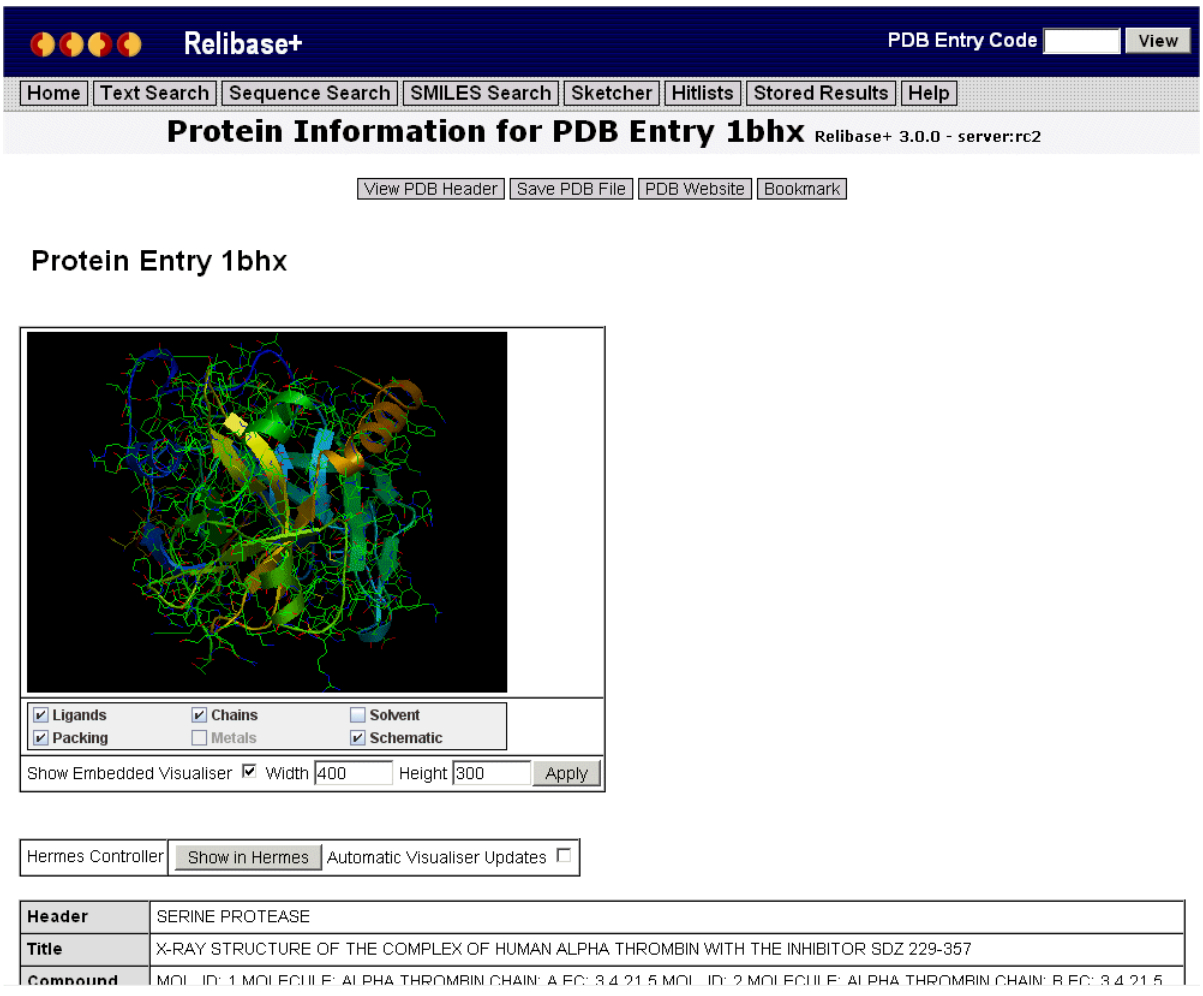
- When you have asked for query atoms or centroids to be superimposed (see Running a Search Section 6.8, page 74), the resulting overlay of hits is displayed in the embedded visualiser window (see AstexViewerTM Section 5.1, page 42). Alternatively, the 3D superposition can be read into Hermes using the **Show in Hermes** button.

4 Viewing Information for Individual Hit Structures

4.1 Relibase+ Protein Entry Pages

Each Relibase+ protein entry page contains:

- Embedded 3D visualisation via AstexViewer (see AstexViewerTM Section 5.1, page 42).
- A Hermes control panel (see Hermes Section 5.2, page 47).
- Protein and ligand information (see Protein and Ligand Information Section 4.3, page 30).
- A link to information on water structure in the entry (see Water Information Section 4.4, page 31).
- A link to information on cavities in the entry (see Cavity Information Section 4.5, page 34).
- A link to information on the secondary structure in the entry (see Secondary Structure Information Section 4.6, page 35).
- Customisable content and hyperlinks to external resources can also be added (see the Relibase+ Installation Notes, Appendix B, *http://www.ccdc.cam.ac.uk/support/documentation/#relibase*).



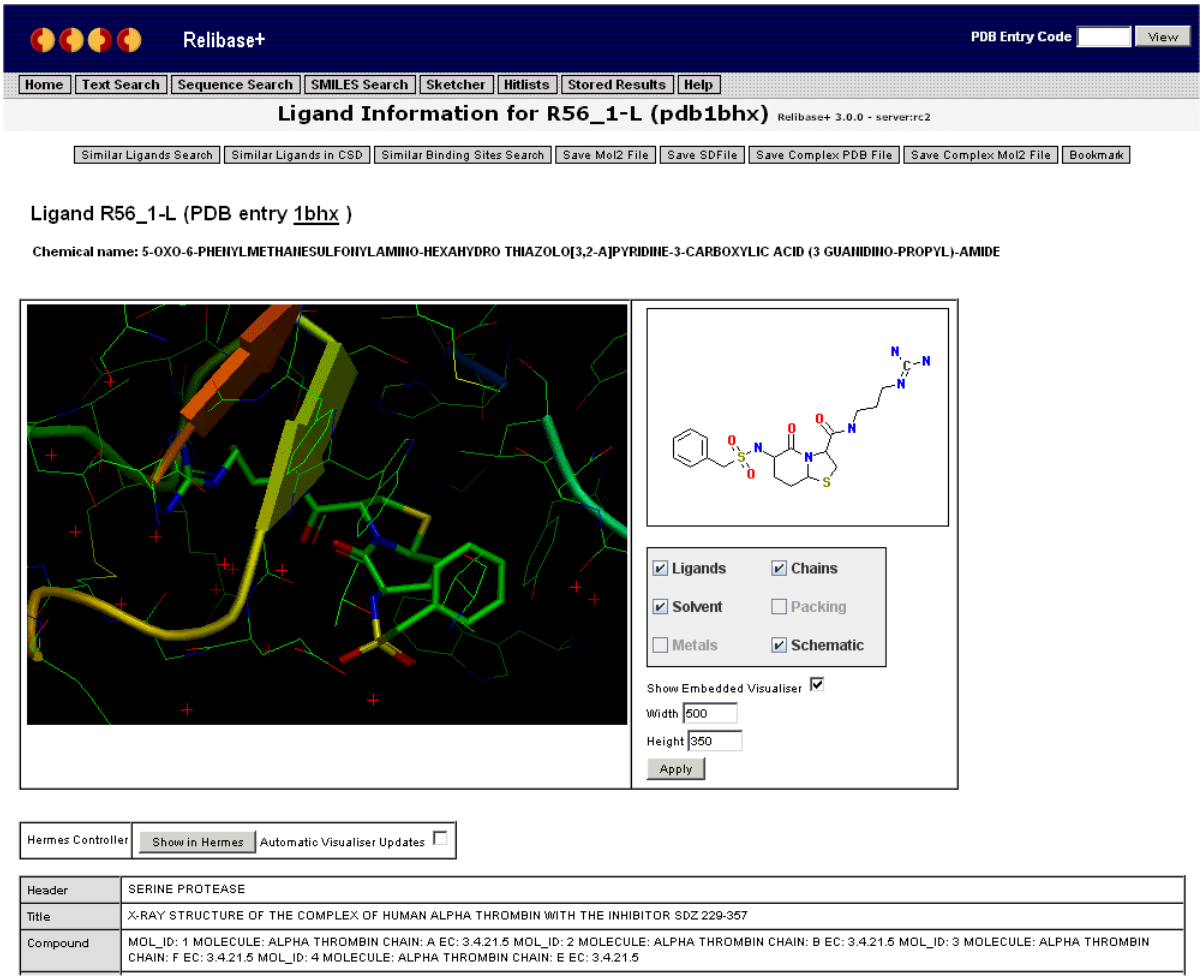
Additionally, the following buttons are present at the top of each protein entry page:

- **View PDB Header:** launches a new browser window with the complete header of the original PDB file.
- **Save PDB File:** export the protein structure in *pdb* file format).
- **PDB Website**: links to the current protein entry on the PDB homepage (*http://www.rcsb.org/pdb/*).
- **Bookmark**: add the current protein entry page to your list of favourites in your browser.

4.2 Relibase+ Ligand Pages

Each Relibase+ ligand page contains:

- Embedded 3D visualisation via AstexViewer (see AstexViewerTM Section 5.1, page 42).
- A Hermes control panel (see Hermes Section 5.2, page 47).
- Protein and ligand information (see Protein and Ligand Information Section 4.3, page 30).
- Information on water structure in the entry (see Water Information Section 4.4, page 31).
- A link to information on cavities in the entry (see Cavity Information Section 4.5, page 34).
- Information on the secondary structure in the entry (see Secondary Structure Information Section 4.6, page 35).
- Customisable content and hyperlinks to external resources can also be added (see the Relibase+ Installation Notes, Appendix B, *http://www.ccdc.cam.ac.uk/support/documentation/#relibase*).

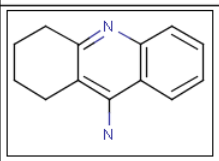


Additionally, the following buttons are present at the top of each ligand page:

- **Similar Ligands Search**: launches a search of the loaded database(s) for ligands similar to the current ligand (see Searching for Similar Ligands in the PDB Section 7.1, page 78).
- **Similar Ligands in CSD**: launches a search of the Cambridge Structural Database (CSD) for ligands similar to the currently ligand (see Searching for Similar Ligands in the CSD Section 7.2, page 80).
- **Similar Binding Sites Search:** launches a search for similar binding sites (see Similar Binding Site Searches (and Superposition) Section 9, page 84).
- **Save Mol2 File**: export the ligand in *mol2* file format.
- **Save SDFile**: export the ligand in *sd* file format.
- **Save Complex PDB File**: export the binding site in *pdb* file format.
- **Save Complex Mol2 File**: export the binding site in *mol2* file format.
- **Bookmark**: add the current page to your list of favourites in your browser.

4.3 Protein and Ligand Information

Protein and ligand information for an acetylcholinesterase complex (protein entry code 1acj) is shown:

Header	HYDROLASE(CARBOXYLIC ESTERASE)
Title	QUATERNARY LIGAND BINDING TO AROMATIC RESIDUES IN THE ACTIVE-SITE GORGE OF ACETYLCHOLINESTERASE
Compound	MOL_ID: 1 MOLECULE: ACETYLCHOLINESTERASE CHAIN: A EC: 3.1.1.7 ENGINEERED: YES
Reference	PROC.NATL.ACAD.SCI.USA **90**, (1993), 9031
Author(s)	J.L.SUSSMAN M.HAREL I.SILMAN
Source	MOL_ID: 1 ORGANISM_SCIENTIFIC: TORPEDO CALIFORNICA
Method	X-RAY DIFFRACTION
Crystal	Cell: a = 113.700, b = 113.700, c = 138.100, alpha = 90.000, beta = 90.000, gamma = 120.000, Spacegroup = P3121
Resolution	2.800
RFactor	19.5%
Deposition Date	18.8.1993
Ligands	 THA in pdb1acj **Water-mediated protein-ligand contact paths: 2**
Protein Chains	pdb1acj-A
Nucleic Acids	No nucleic acids for this entry.
Secondary Structure	Secondary Structure Information
Waters	Water Information
Cavities	Cavity Information

For a typical entry the following information is given:

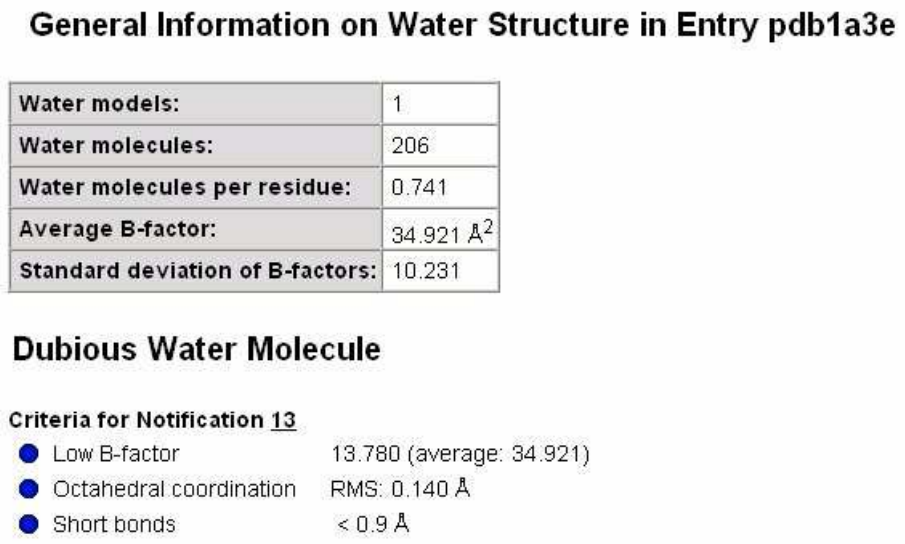
- A summary of the textual, bibliographic and crystallographic information for this protein entry, including **Header, Title, Compound, Reference, Author(s), Source, Method, Crystal, Resolution, RFactor,** and **Deposition Date**. Click on any author's name to run an author search (see Author Name Searches Section 3.2, page 59) on that name.
- A *Caveat* record is present for PDB entries that contain information under the *CAVEAT* section (i.e. are considered to be in error by the PDB) in their PDB file.
- In the example above, the structure contains only one ligand (and binding site), a tacrine molecule. Clicking on the 2D chemical diagram of the tacrine molecule will link to the ligand page (see Relibase+ Ligand Pages Section 4.2, page 28) for this ligand/binding site.
- The amino acid chains in the structure are listed; in the example above there is only one chain. The protein chain sequence can be viewed by clicking on the **Chain Identifier** hyperlink. Searches for similar chains (see Similar Protein Chain Searches Section 8, page 81) can also be initiated from the resulting page.
- Information on water structure in the entry can be accessed by clicking on the **Water Information** button (see Water Information Section 4.4, page 31).
- Information on cavities in the entry can be accessed by clicking on the **Cavity Information** button (see Cavity Information Section 4.5, page 34).

- Information on the secondary structure in the protein can be accessed by clicking on the Secondary Structure Information button (see Secondary Structure Information Section 4.6, page 35).
- Customisable content and hyperlinks to external resources can also be added (see the Relibase+ Installation Notes, Appendix B, *http://www.ccdc.cam.ac.uk/support/documentation/#relibase*).

4.4 Water Information

4.4.1 Information on Water Structure

- Information about the water structure for each database entry is accessed by clicking on the **Water Information** button at the bottom of either the Relibase+ protein entry page (see Relibase+ Protein Entry Pages Section 4.1, page 27) or the Relibase+ ligand page (see Relibase+ Ligand Pages Section 4.2, page 28):



General Information on Water Structure in Entry pdb1a3e

Water models:	1
Water molecules:	206
Water molecules per residue:	0.741
Average B-factor:	34.921 Å^2
Standard deviation of B-factors:	10.231

Dubious Water Molecule

Criteria for Notification 13

Low B-factor	13.780 (average: 34.921)
Octahedral coordination	RMS: 0.140 Å
Short bonds	< 0.9 Å

- General information on the water structure is given in the above table, along with the criteria for notification of dubious water molecules (see Data-Related Issues Section 5.6.3, page 8).
- Information about water clusters and rings are also given. The cluster or ring size is displayed along with the individual water molecules which form the structure.

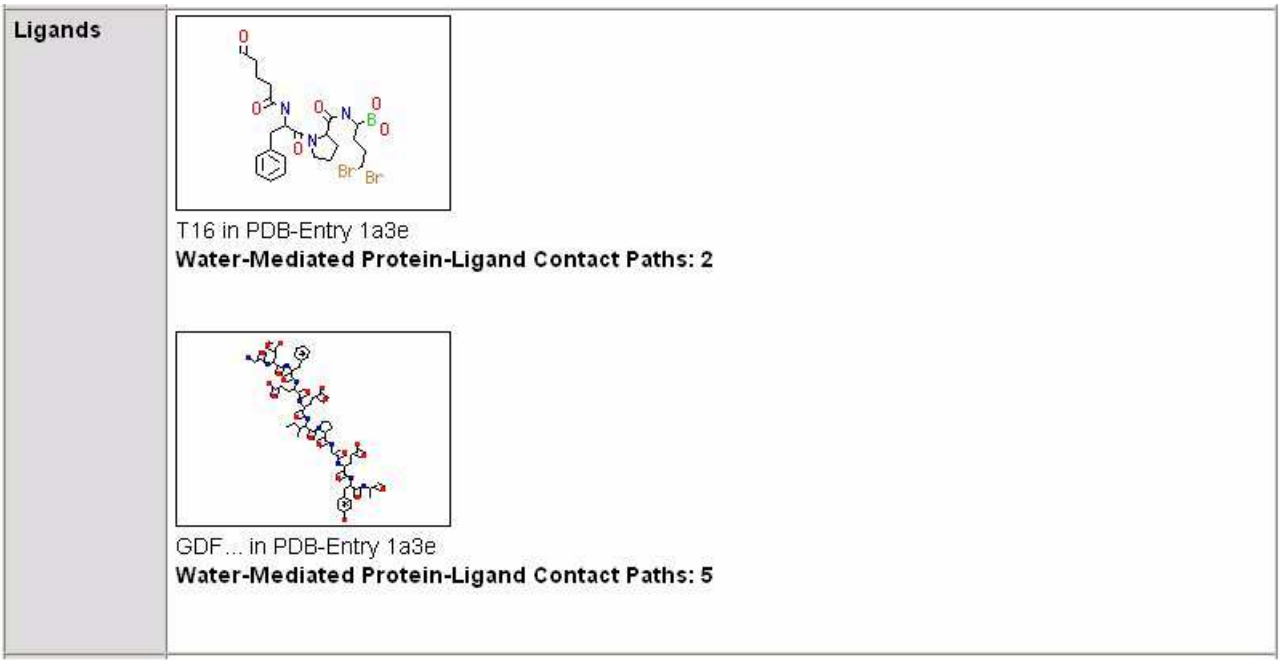
Water Clusters in Entry pdb1a3e

Size	Show	Hermes														
6	☐	Show	2	7	15	33	44	25								
14	☐	Show	24	13	40	14	75	68	88	62	37	12	57	91	121	131
3	☐	Show	163	19	206											
3	☐	Show	6	23	35											
4	☐	Show	119	27	199	69										
4	☐	Show	32	28	211	171										
5	☐	Show	117	31	120	51	99									
6	☐	Show	39	34	148	157	155	168								
3	☐	Show	8	36	55											
6	☐	Show	110	49	151	174	143	147								
5	☐	Show	20	50	129	105	139									
6	☐	Show	90	54	95	136	133	87								
4	☐	Show	61	65	98	164										
3	☐	Show	103	66	128											
3	☐	Show	176	109	190											
3	☐	Show	38	113	41											
3	☐	Show	9	166	183											

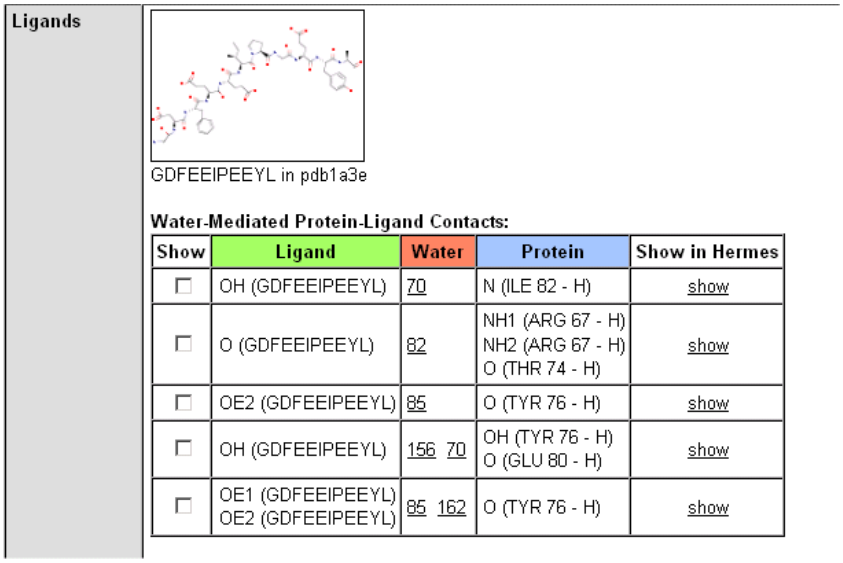
- Clicking on any of the hyperlinkable water molecules in the structure will lead you through to water descriptor information (see Water Molecule Descriptors Section 5.6, page 4).
- To view certain water clusters or rings in the embedded visualiser, activate the appropriate tickbox underneath the column headed with the **Show** button. All clusters/rings can be viewed or hidden by hitting the **Show** button.
- To view certain water clusters or rings in Hermes, click on the **Show** hyperlink adjacent to the appropriate cluster/ring in the table, after ensuring that the full complex structure has been first loaded into Hermes via the **View in Hermes** button.

4.4.2 Water-Mediated Protein-Ligand Contacts

- Information concerning the number of water-mediated protein-ligand contact paths is given under the hyperlinkable 2D ligand diagram on the protein entry page:



- Clicking on the 2D ligand diagram of e.g. the ligand with 5 water-mediated protein-ligand contact paths provides the following information on the ligand information page:

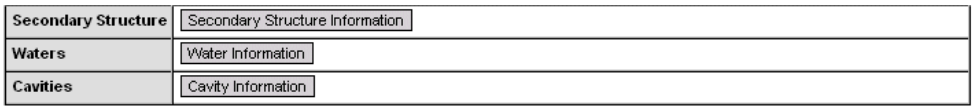


Show	Ligand	Water	Protein	Show in Hermes
☐	OH (GDFEEIPEEYL)	70	N (ILE 82 - H)	show
☐	O (GDFEEIPEEYL)	82	NH1 (ARG 67 - H) NH2 (ARG 67 - H) O (THR 74 - H)	show
☐	OE2 (GDFEEIPEEYL)	85	O (TYR 76 - H)	show
☐	OH (GDFEEIPEEYL)	156 70	OH (TYR 76 - H) O (GLU 80 - H)	show
☐	OE1 (GDFEEIPEEYL) OE2 (GDFEEIPEEYL)	85 162	O (TYR 76 - H)	show

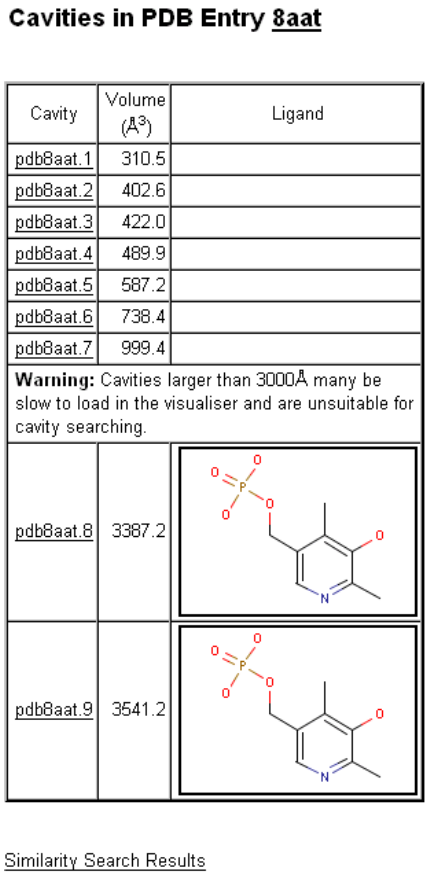
- Clicking on any of the hyperlinkable mediating water molecules, will lead you through to water descriptor information (see Water Molecule Descriptors Section 5.6, page 4).
- In order to highlight certain water-mediated protein-ligand contacts in the embedded visualiser, activate the relevant tick box under the column headed with the **Show** button. Click on the relevant **Show** hyperlink under the column headed *Shown in Hermes* to view the water mediated protein-ligand contacts in Hermes.

4.5 Cavity Information

- Cavity information for any database entry can be accessed by clicking on the **Cavity Information** button at the bottom of either the protein entry or ligand pages:



- If you are on a ligand page and there is information available for the ligand cavity, clicking on this button takes you to a page displaying the volume of the cavity containing the ligand, header information, and the ligand chemical diagram (if available). Also, the visualiser (see Displaying and Comparing Cavities Section 3, page 101) will open with the selected cavity loaded.
- If you are on a protein entry page, clicking on the **Cavity Information** button takes you to a page listing all the cavities in the protein structure, sorted in ascending order of cavity size, and any ligands they contain (clicking on a ligand diagram links you to the *Ligand Information* page):



- Selecting one of these cavities will then give you further details of that cavity and load it into the visualiser (see Displaying and Comparing Cavities Section 3, page 101).

4.6 Secondary Structure Information

- The term secondary structure describes the general 3D form of local segments of the protein (i.e. amino acids). Secondary structure in proteins is typically mediated by H-bonding interactions between amino acids.
- The methodology involved in the display of secondary structure in Relibase+ is covered in the sections that follow.

4.6.1 Introduction

- Secondary structure is assigned to protein structures to derive a sense of the relative fold of one protein with respect to another. There are many examples of assignment protocols published in the literature. The most widely used method is known as *Define Secondary Structure of Proteins* (or DSSP, Kabsch & Sanders) but others also exist.
- Secondary structure assignments usually operate by recognising particular intra-molecular non-bonded features between given residues in a protein. Pairs of residues that exhibit these pre-defined features are then assigned as being a component of a helix or sheet. Strands and turns are also observed, and can be sub-components of helices or sheets.
- In raw PDB entries, information is presented about helices and sheets. Some turn information is also presented, but the data are on average usually limited to a single turn per chain. This does not reflect the full secondary structure assignment in a given PDB entry, as a given chain can contain several isolated turns that are not sub-components of a helix.
- A more serious issue with secondary structure classification in the PDB, for the purposes of database searching, is that the classification method used varies from structure to structure. This means that the definition of an N-terminus of a helix (for example) will differ in one PDB entry to the next, according to the definition included with the public structure.
- The secondary structure module in Relibase+ contains the original PDB assignments of secondary structure but, in addition, contains a new consistent assignment of turns and helices. The turn assignment is based on a machine learning algorithm to cluster various different turn types and was used for a complete assignment of all turns in given proteins. This turn assignment was also used to define helices based on multiple turns. Furthermore, algorithm for the identification of kinks and bends in helices and b-strands are implemented.

4.6.2 Classification Methodology

- A basic description of the methodology used is given here. For a more complete description please refer to the following references:
  - Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions,
    O. Koch, G. Klebe, *Proteins: Structure, Function, and Bioinformatics*, 74, 353-367, 2008.
    [DOI: 10.1002/prot.22185]

A full description of all turn types is available as supplementary material for the above reference.

- *http://archiv.ub.uni-marburg.de/diss/z2008/0730/*. This reference also includes the amino acid propensities for all turn types.
- Publications on SHAFT and the secondary structure module in Relibase+ are currently in preparation.

- Secondary structure assignments are classified into the following:
  - **Turn Assignment** (see page 36)
  - **Turn Types** (see page 36)
  - **SHAFT Assignment of Helices** (see page 37)

4.6.3 Turn Assignment

- A diverse subset of 1903 chains was used as a training set for assigning of turn types. Initially, turns were identified in the training set.
- Each sequence of up to 6 residues in each peptide chain was extracted and analysed to identify close contacts between the terminal residues: Several types of contact were identified: Hydrogen bonds were deemed to be present if the DSSP function (Kabsch and Sander, in-house implementation) indicated their presence. Hydrogen-bound turns were then classified as either 'normal' or 'reverse' dependent on the direction of the hydrogen bond with respect to the direction of the peptidic sequence. Additionally, 'open' turns were identified, where the Cα…Cα distance between the terminal residues in the sequence was less than 10Å. Sequences that fell into one of these classifications was deemed to be a 'turn'. The subdivision nominally could results in 15 subset, namely normal, inverse and open turns of 2,3,4,5 and 6 residues respectively. In practice, 3 normal turn families, 4 open turn families and 5 reverse turn families were observed, and clustered.
- For each turn, back-bone torsion angles were evaluated and used as coefficients of an N-dimensional vector (Where N was the number of backbone torsion angles in each turn type). The vectors were then used for clustering based on Euclidian distance, by making each observed sequence vector a node in an Emergent Self-Organising Map (ESOM). The ESOM clusters similar vectors into similar regions, leading to a self organised classification of turns for structures in the training set.
- Analysis of the results led to identification of 158 turn types, each of which is identified in Relibase+. Using the trained ESOMs, the turn types can then be assigned programatically to all entries in the PDB, leading to a consistent turn type assignment across all entries.

4.6.4 Turn Types

- The resulting classification of secondary structure enumerates many different turn types. Each have been given a classification based on residue length, turn 'type' (open, reverse or normal)

and a sub-classification based on the cluster occupied by a given turn type. These assignment names are based on the ranges of internal backbone angles within a give turn type. There is no firm nomenclature, but turns that are similar in nature tend to have assignment names that are similar.

- For example - below is the set mean torsion angles for d-turns (reverse 3-residue turns): types Ia and Ib are relatively similar - the major difference is in ψ_2; all other torsion angles have overlapping ranges.

cluster	type	ϕ_1		ψ_1		ω		ϕ_2		ψ_2	
		f	+/-	f	+/-	f	+/-	f	+/-	f	+/-
1	Ia	124.0	21.0	-51.0	27.3	175.9	6.4	126.7	34.2	155.1	15.5
2	Ib	116.2	17.7	49.8	18.2	174.4	7.4	148.6	22.7	169.5	12.6
3	IIa	127.4	21.6	-26.3	22.2	176.4	6.8	-87.0	29.4	148.5	24.9
4	IIb	104.8	10.0	-24.5	25.6	174.1	4.6	-105.2	8.7	-172.8	3.6
5	IIIa	-113.2	19.6	27.4	30.1	-177.5	5.3	124.2	32.7	162.9	15.1
6	IIIb	-131.8	21.6	42.2	35.5	-175.0	6.2	84.5	30.8	-160.8	15.8
7	IIIc	166.2	11.1	32.7	23.9	-176.5	0.1	127.5	30.0	-163.9	0.2
8	IVa	-148.4	13.8	98.2	12.4	-0.4	3.9	-74.0	12.6	159.7	15.3
9	IVb	131.7	23.4	115.5	7.9	-9.7	13.3	-57.9	9	173.1	5.1

- In cases where a given turn sub-classification resultant from ESOM clustering corresponds to a previously defined sub-classification, the name used corresponds to that in the literature. So, for example, g-turns (normal 3-residue turns) are classified into 'inverse' or 'normal' sub-classifications.
- A full description of all turn types is available in the publications provided in the **Classification Methodology** (see page 35).

4.6.5 SHAFT Assignment of Helices

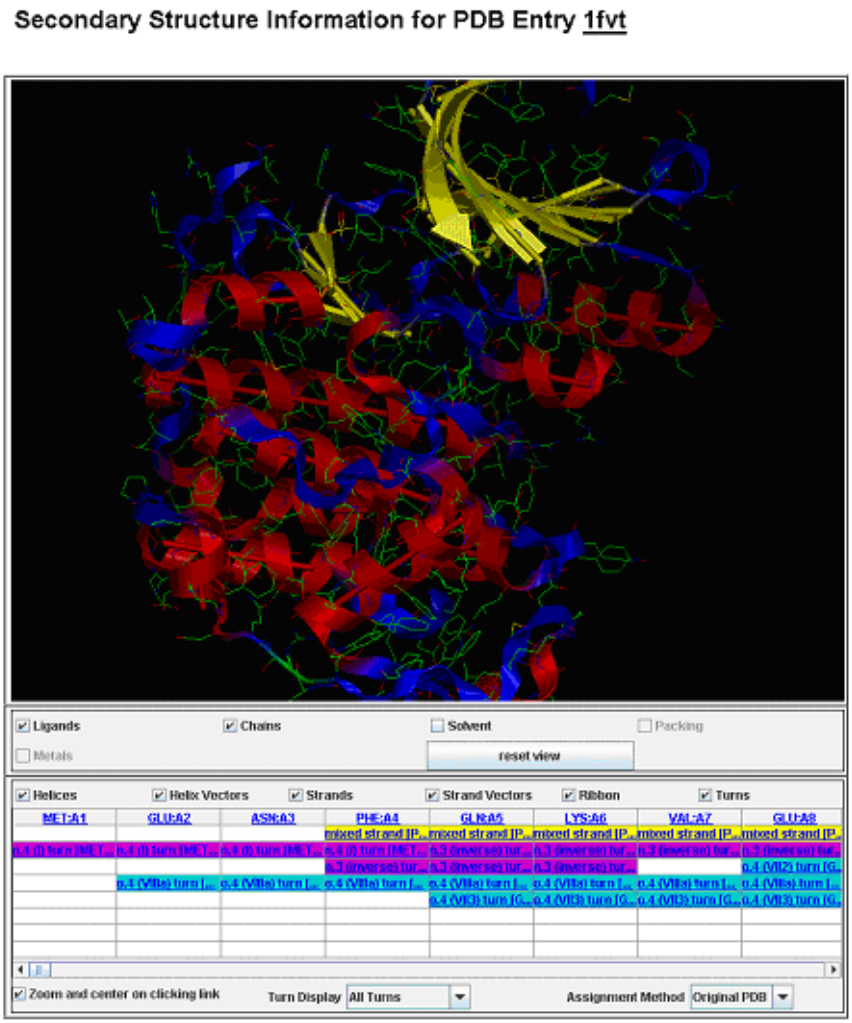
- Once turns have been assigned, it is relatively straight-forward to build up a given sequence of turns into a larger secondary structure element.
- The general feature of a-helices is the intra-helical NH_i - CO_i+4 hydrogen bonding, so that each residue within a helix is involved in two backbone hydrogen bonding (NH_i - CO_i+4 and CO_i - NH_i-4). The first and last four residues within an a-helix are only involved in one type of backbone hydrogen bonding, leaving four NH-groups at the N-terminus and four CO groups at the C-terminus that can interact with other partners, e.g. other parts of the protein or water.

Therefore, in an ideal a-helix the N3 position would be the last residue with this "free" backbone NH at the N-terminus and the C3 position would be the first residue with a "free" backbone CO group at the C-terminus. The assignment of the N-capping and C-capping position using turn motifs is thus based on this "free" backbone groups and the analysis of specific turn types at the terminus.

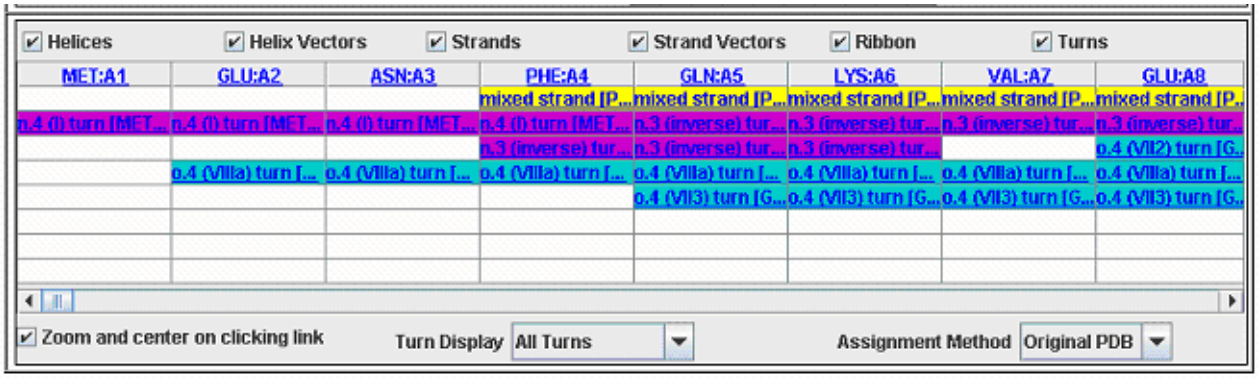
- This approach was used to convert sequences of turns into helical elements. If one recognises the turn types that occur in helices, then one can build up sequences of contiguous turns that are of the appropriate type and assign them to a specified helix type.
- Helix capping residues can be assigned as those that are at the ends of the contiguous sequences. Some issues with overlapping helices of different types were identified where, for example, an a-helix would be overlapped with a 3_{10}-helix. This was resolved by coalescing the overlapping helices and re-assigning the coalesced helix to a given type via a set of hierarchical rules. Fuller details of the methodology used are available, see the URL above for further information.

4.6.6 Viewing Secondary Structure Assignments

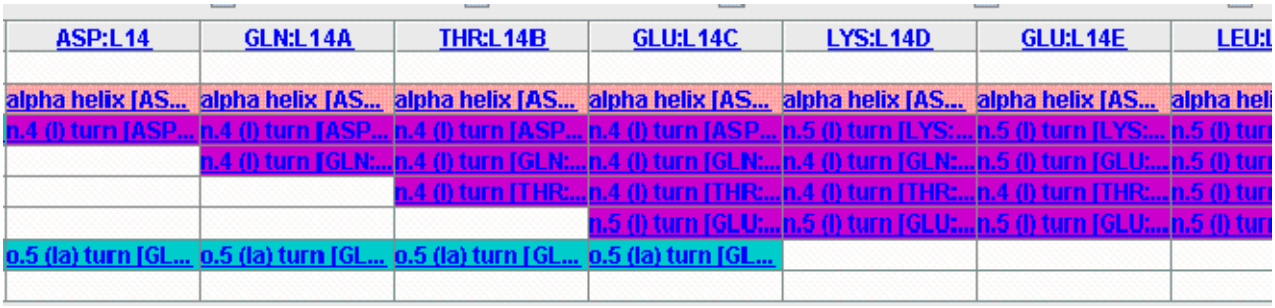
- Information about secondary structure in a protein is accessed via the **Secondary Structure Information** button at the bottom of protein or ligand information pages.
  Note: secondary structure shown in the main pages in Relibase+ is that as assigned by AstexViewer using the DSSP algorithm, rather than the assignment from the secondary structure module.
- After clicking on the **Secondary Structure Information** button, a *Secondary Structure Information* page is opened that illustrates the secondary structure assigned to the selected entry. An example page is shown below for entry 1fvt.



- The page contains an instance of AstexViewer containing the view of a given protein. Beneath the AstexViewer display are some controls for managing what is displayed. Use the various tick boxes (*Ligands*, *Chains*, *Solvent*, *Packing* and *Metals*) to switch the named component on and off. Use the **reset view** button to return the display back to its initial view.
- Also provided is a secondary structure browser that permits navigation of secondary structure elements in the protein:



- Each cell in the table corresponds to a secondary structure assignment to a given amino acid in the protein. Cells contain a condensed description of the secondary structure element in question. For turns the convention used is x.y (sub type), for example, an inverse gamma turn would be described as `n.3 (inverse)` in the table.
- Hovering over a cell with the mouse cursor displays the full description of the element in question. Turns, helices and strands are coloured differently to aid recognition. Turns are coloured differently based on type (normal are magenta, open are cyan and reverse are green) and then shaded by length.
- A given amino acid can be a component in one or more secondary structure elements: this particularly applies to turns as single amino acids can be part of several overlapping turns that build up to make a compound secondary structure element.
- The most common example of such behaviour is well known: an ideal `a`-helix in a protein can be viewed as a sequence of 5-residue type 1 turns that overlap with each other. Below is an example of an `a`-helix that was assigned using SHAFT starting with ‘normal’ 4-residue turns:

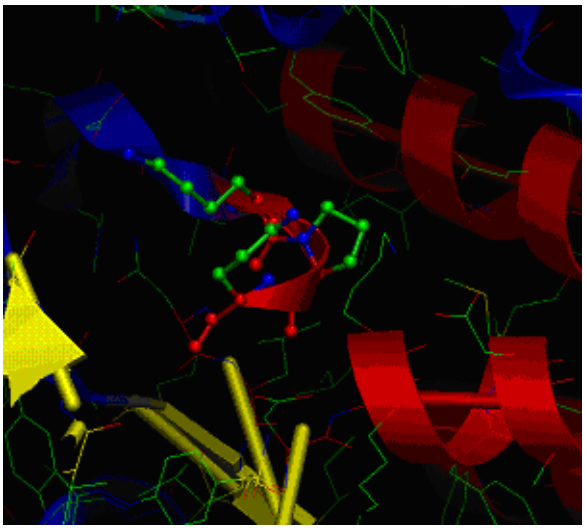


- Beneath the table of secondary structure assignments there are a number of other options:
  - **Zoom and center on clicking link**: keep this tick box checked if you wish to highlight, centre and zoom the selected secondary structure element; uncheck the tick box if you wish to only highlighted the secondary structure element (the state of the centring and zoom will be unaffected).

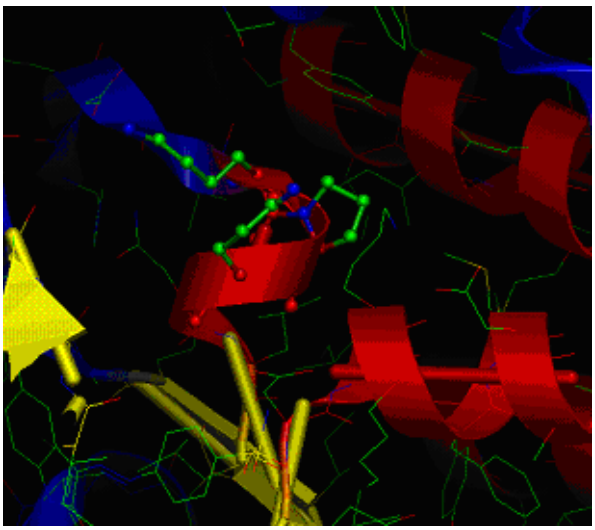
- *Turn Display*: select options in this pull down menu to control whether to display *All Turns*, *Rare Turns* (all turns except those that are parts of a-helices), or *No Turns*.
- *Assignment Method*: use this pull down menu allows to alternate between helices as pre-assigned in the original PDB file compared to those assigned using the SHAFT methodology. Turn information is always presented using results from the SecBase assignment.
- 2D/3D searches can be constrained so that the resultant hits contain specific secondary structure elements (see Defining Secondary Structure Elements Section 13, page 146).

4.6.7 Helix assignments from the PDB versus SHAFT assignment: notable differences

- Using the SHAFT for building helices, for the most part, results in assignments that are broadly similar to PDB assignments. The major differences so far noted are in the termini of helices. Further, SHAFT assigns more a-helices.
- Helical Termini:
  - Generally, SHAFT tends to extend helices to include additional residues, most frequently at the C-terminus end. Visual inspection suggests that these alterations generally make sense. For example, in PDB entry 1fvt, glutamine A131 is regarded as the C-terminus of a 3_{10} helix. SHAFT extends the 3_{10} helix to include residues ASN A132 and LEU A133.
  - Images of the 2 assignments can be seen below. In the image, one can also see that the SHAFT assignment has extended an *a*-helix by one residue: This is also visible in the image below.



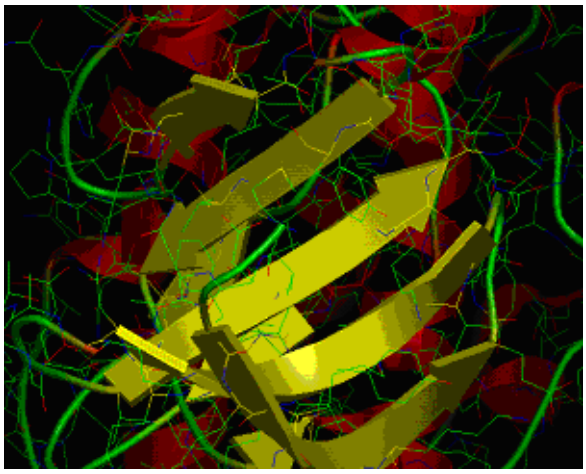
Original PDB Assignment



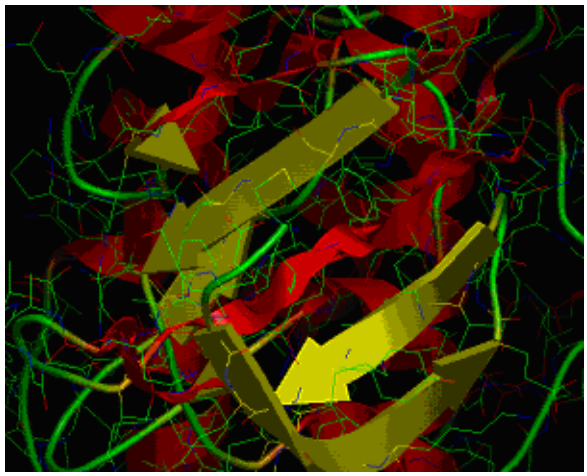
SHAFT Assignment

- Gamma Helices:
  - SHAFT can assign *g*-helix status to regions previously deemed as parts of sheets which is well known in the literature. This assignment is essentially subjective - in practice, the gamma helical status of the sheet strand is additional to its assignment as part of a sheet rather than an alternative, as the gamma helix may still make the designated interactions that one associates

with a sheet. An example of this behaviour can be seen in 1fvt: A strand in the centre of a twisted beta-sheet is reassigned as a gamma helix.



Original PDB Assignment



SHAFT Assignment

5 3D Visualisation of Structures

- The two visualisers provided with Relibase+ serve slightly different purposes. AstexViewer (see AstexViewerTM Section 5.1, page 42) is embedded in the Relibase+ interface to provide quick and easy visualisation of hit structures, including the display of multiple structures. Hermes (see Hermes Section 5.2, page 47) is provided to facilitate more detailed investigation of the hit structures.

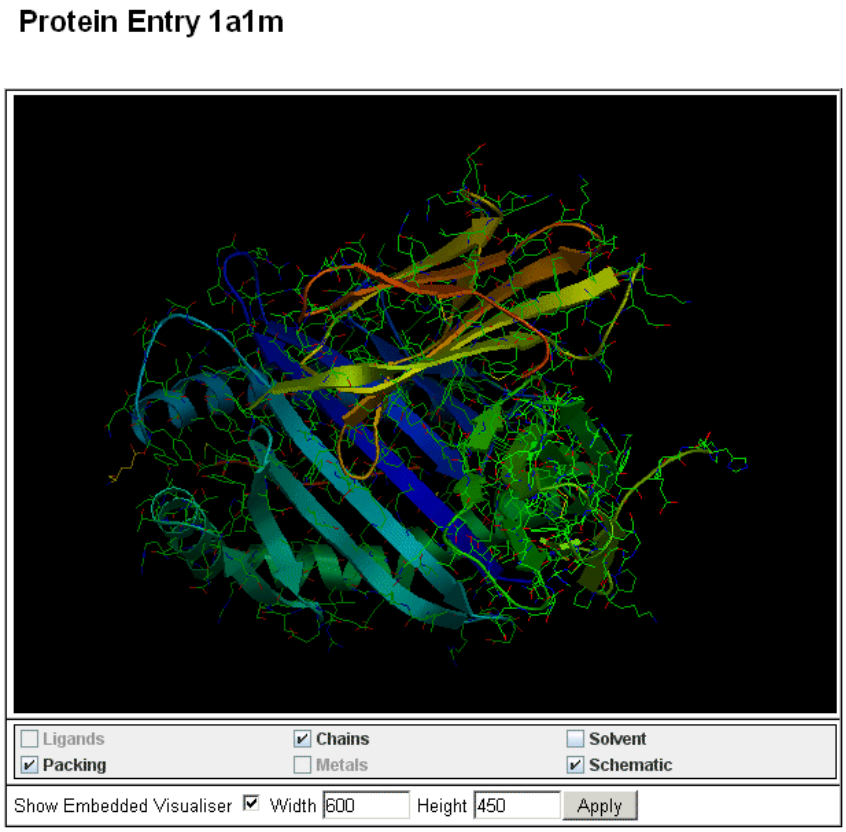
5.1 AstexViewer™

- AstexViewer is a Java molecular graphics program developed and distributed by Astex Therapeutics: *http://www.astex-therapeutics.com*.
- Basic functionality is documented below. Advanced functionality is available by right-clicking. More detailed information can be found in the AstexViewer documentation: *http://www.astex-therapeutics.com/AstexViewer/AstexViewer2/doc/interface.html*. This documentation is also provided in the Relibase+ distribution.
  - **Rotation**: move the cursor around in the 3D window while keeping the left-hand mouse button pressed down.
  - **Translation**: move the cursor left or right in the 3D window while keeping the left hand mouse button and the Control key pressed down.
  - **Scale**: zoom in or out by moving the cursor up and down in the in the 3D window while keeping the left-hand mouse button *and* the Shift key pressed down.
- The appearance of the 3D display varies depending on the type of page.
  - Protein information page (see Appearance of AstexViewer on a Protein Information Page

Section 5.1.1, page 43).

- Ligand information page (see Appearance of AstexViewer on a Ligand Information Page Section 5.1.2, page 44).
- Binding site information page (see Appearance of AstexViewer on a Binding Site Superposition Page Section 5.1.3, page 45).
- Secondary structure display (see Secondary Structure Display Section 5.1.4, page 46).

5.1.1 Appearance of AstexViewer on a Protein Information Page

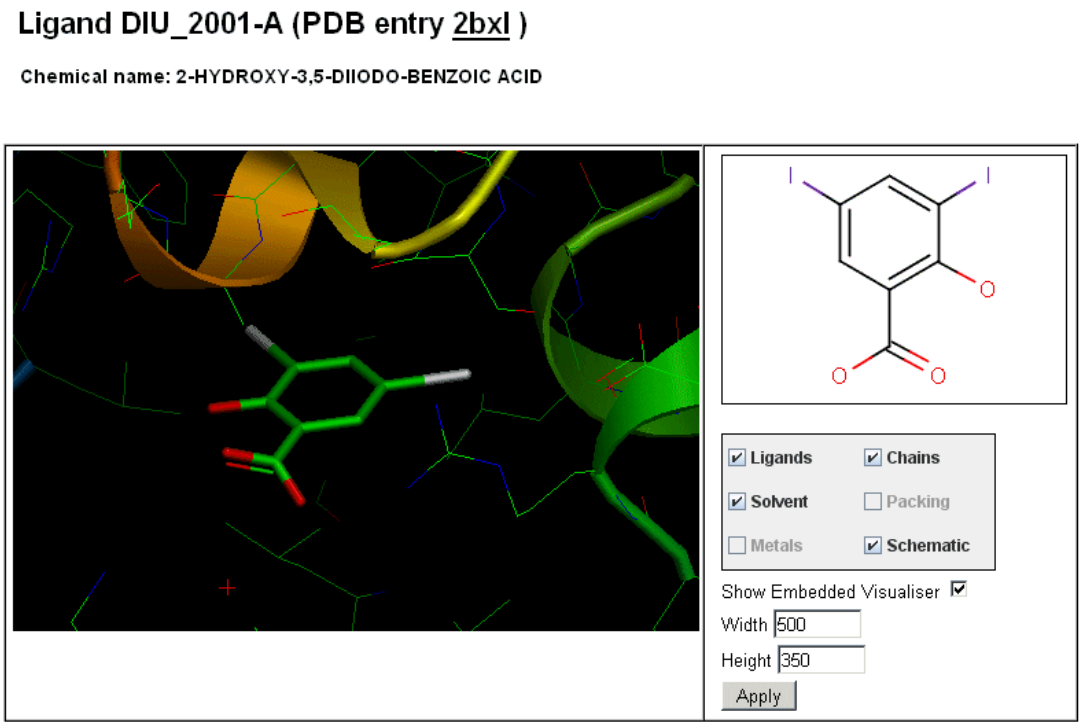


- The display can be rotated, translated and scaled as previously described (see AstexViewerTM Section 5.1, page 42).
- A number of check boxes beneath the display can be used to control the view:
  - **Ligands**: displays or hides ligands.
  - **Packing**: displays or hides any crystal packing present in the protein.
  - **Chains**: displays or hides protein chains.
  - **Metals**: displays or hides metal atoms.
  - **Solvent**: displays or hides solvent molecules.
  - **Schematic**: displays or hides the protein cartoon display.

Note: the schematic is assigned by AstexViewer and not Relibase+ nor the PDB.

- Whether or not AstexViewer is displayed is controlled via the *Show Embedded Visualiser* tickbox. The dimensions of the viewer size (default is 800 by 600) may also be modified to fit user requirements by typing new dimensions into the *Width* and *Height* windows then hitting **Apply**.

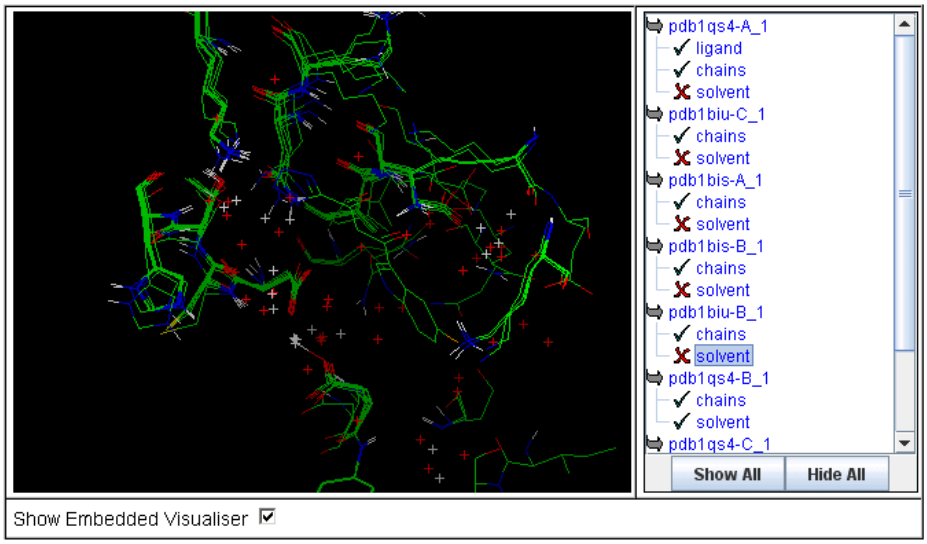
5.1.2 Appearance of AstexViewer on a Ligand Information Page



- In addition to the 3D display, a 2D diagram is also provided.
- The display can be rotated, translated and scaled as previously described (see AstexViewerTM Section 5.1, page 42).
- A number of check boxes beneath the display can be used to control the view:
  - **Ligands**: displays or hides ligands.
  - **Solvent**: displays or hides solvent molecules.
  - **Metals**: displays or hides metal atoms.
  - **Chains**: displays or hides protein chains.
  - **Packing**: displays or hides any crystal packing present in the protein.
  - **Schematic**: displays or hides the protein cartoon display
    Note: the schematic is assigned by AstexViewer and not Relibase+ nor the PDB.
- Whether or not AstexViewer is displayed is controlled via the *Show Embedded Visualiser*

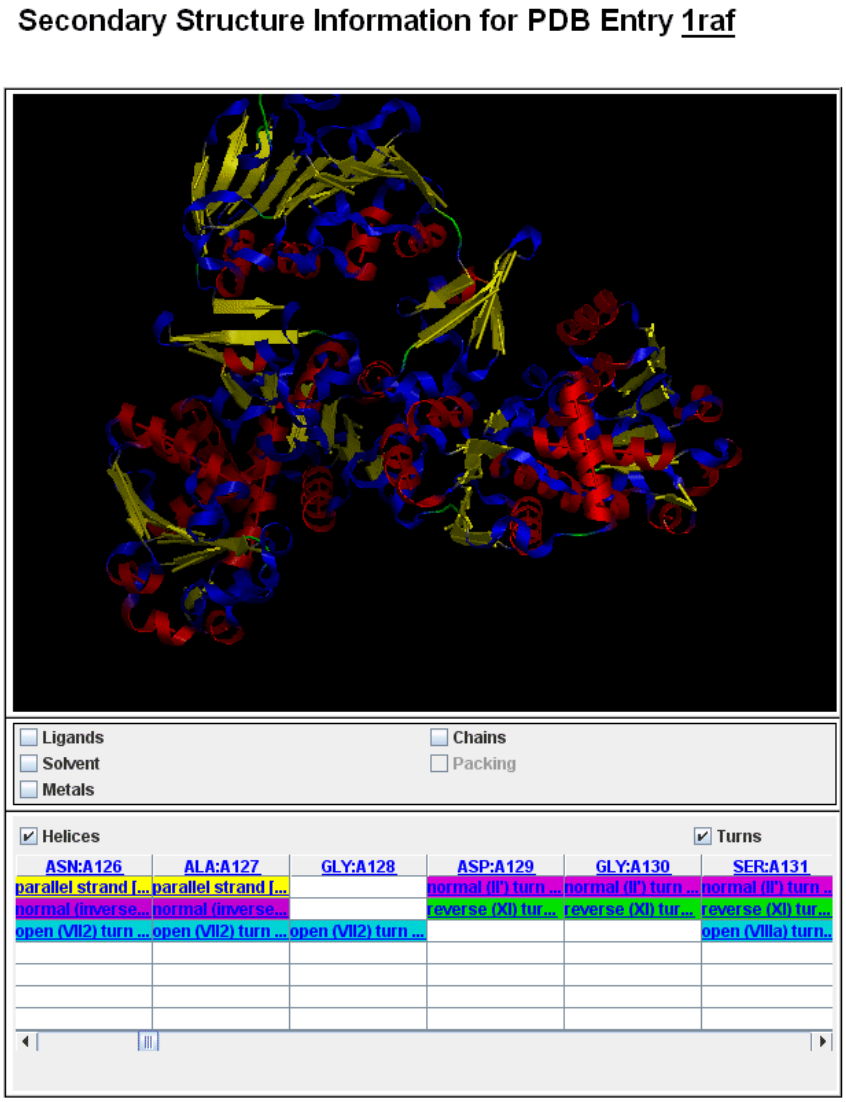
tickbox. The dimensions of the viewer size (default is 500 by 350) may also be modified to fit user requirements by typing new dimensions into the *Width* and *Height* windows then hitting **Apply**.

5.1.3 Appearance of AstexViewer on a Binding Site Superposition Page



- The 3D display contains the superimposed binding sites while the pane to the right controls the view in the 3D display.
  - An entire PDB entry (including protein chains, ligands and solvent) can be switched on or off by clicking on the grey arrow adjacent to the PDB code, e.g. *pdb1qs4-A_1* in the screenshot above.
  - The display of components of the PDB entry (protein chains, ligands and solvent) can be controlled by clicking on the relevant word, adjacent to the green tick or red cross. In the case of *pdb1k6y-B_1* above, the protein chains are displayed but the solvent is not.
  - Use the **Show All** and **Hide All** buttons to control the global display.
  - Control whether or not AstexViewer is displayed via the *Show Embedded Visualiser* tickbox.

5.1.4 Secondary Structure Display



- The display can be rotated, translated and scaled as previously described (see AstexViewerTM Section 5.1, page 42).
- A number of check boxes beneath the display can be used to control the view:
  - **Ligands**: displays or hides ligands.
  - **Solvent**: displays or hides solvent molecules.
  - **Packing**: displays or hides any crystal packing present in the protein.
  - **Chains**: displays or hides protein chains.
  - **Metals**: displays or hides metal atoms.
- Further details on controlling the display of secondary structure elements such as helices etc are provided elsewhere (see Secondary Structure Information Section 4.6, page 35).

5.2 Hermes

- Hermes is a program for visualising protein structures in three dimensions, with particular emphasis on functionality for the analysis of protein-ligand binding interactions.
- Hermes can be launched from any protein entry or ligand page by hitting the **Show in Hermes** button within the *Hermes Controller* section of the interface:

- To have the visualiser update automatically to display the structure currently shown in the browser switch on the **Automatic Visualiser Updates** check box. If this is switched off the current structure will remain in the display until the **Show in Hermes** button is clicked again.
- Use of Hermes is covered in detail elsewhere (follow the Hermes documentation link on the right of this page).

6 Storing, Combining and Converting Search Results

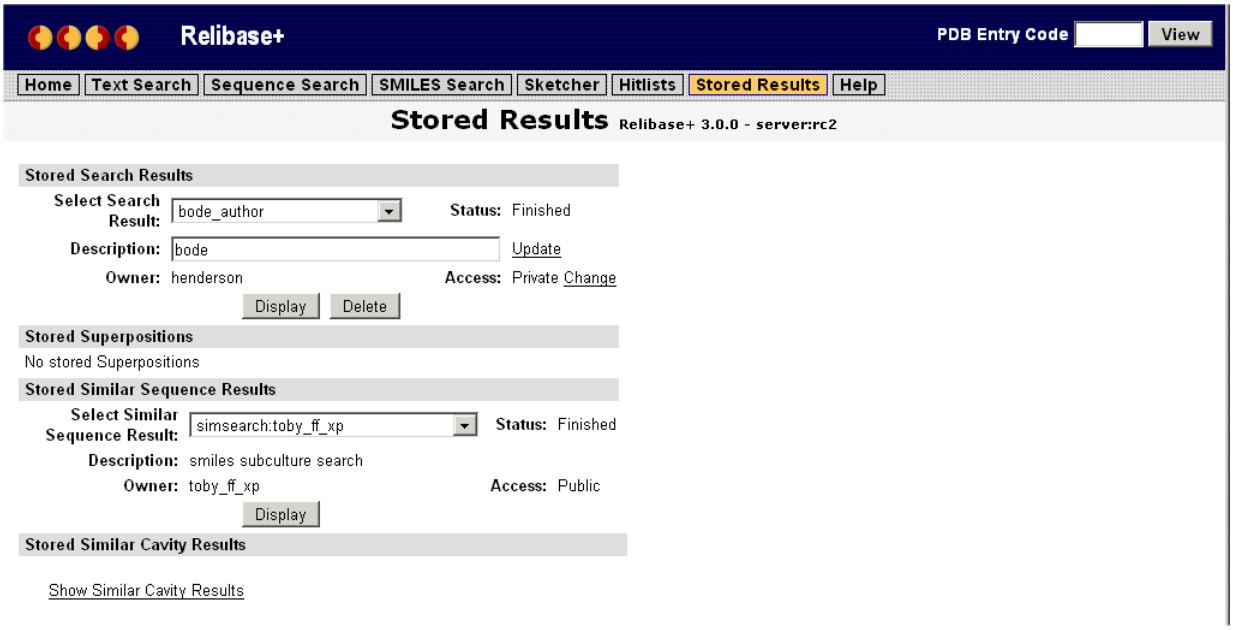
- Search results can be saved in one of two ways depending on which type of search has been run:
  - Storing of search results (see Storing Search Results Section 6.1, page 47).
  - Saving of search results in hitlists. Hitlists are lists of entries saved from Relibase+ searches, stored separately for each Relibase+ user. Relibase+ uses three types of hitlist (protein, ligand, and cavity).

6.1 Storing Search Results

- Results from text searches, similar binding site superpositions, sequence searches and cavity similarity searches can be stored. This is done by typing a search name into the *Save Superposition Results* or *Save Sequence Results* part of the window (generally found at the bottom of the page) as well as a description of the search, then hit **Save**.

- For text searches hit the **Save Search Results** button on the bottom right of the page, enter a name for the search results then hit **Save**.

- Stored searches (i.e. for searches run in batch mode (see Options available on Search: Filters Section 6.8.1, page 75) can be accessed from the *Stored Results* window. Cavity similarity searches are saved automatically and can also be viewed:



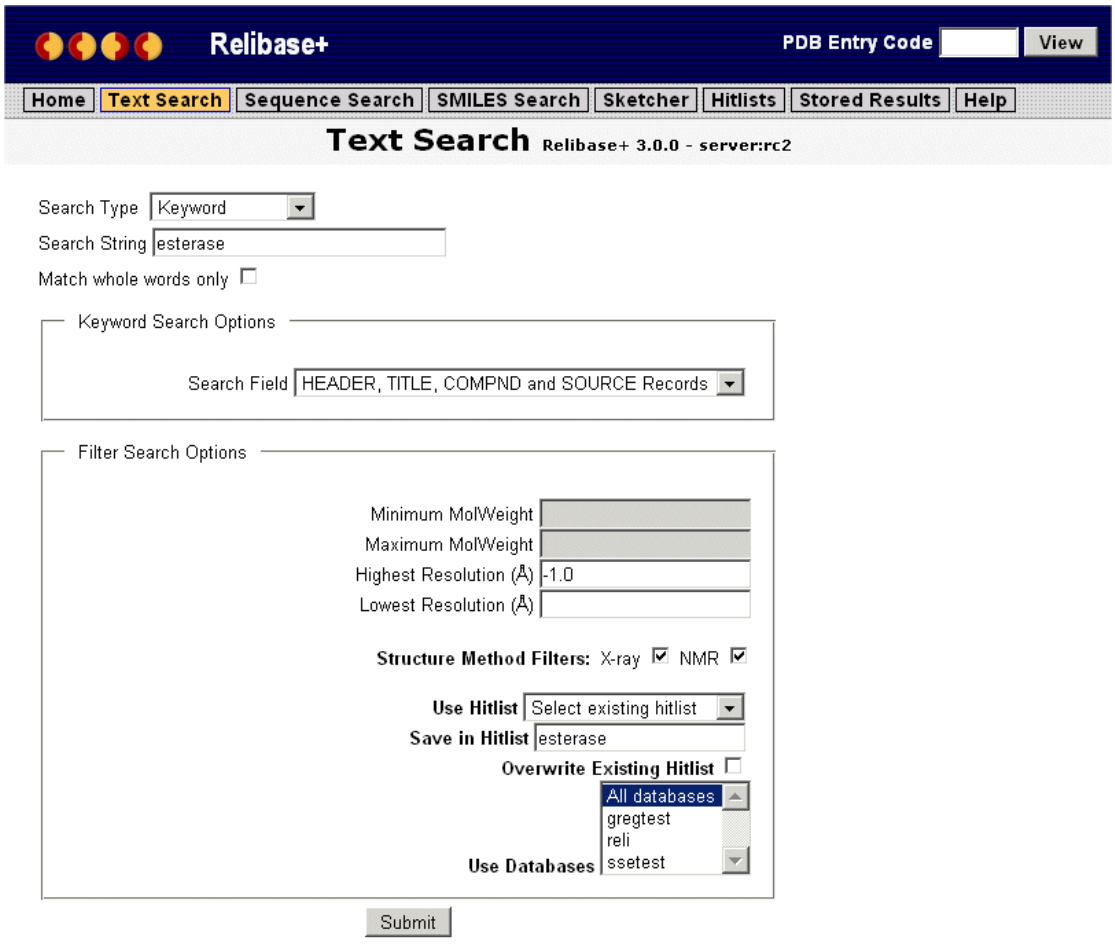
- It is not possible to edit, combine or manage stored search results.

6.2 Creating Hitlists

- Search results from the following types of searches can be stored as hitlists:
  - **Text** searches (see Keyword Searches Section 3, page 58). Results from this type of search can be selected to be saved either ***before*** or ***after*** the search has been run.
  - **Sequence** searches (see Protein Sequence Searches Section 4, page 65). Results from this type of search can only be selected to be saved ***before*** the search has been run. In addition to being initiated from the Relibase+ menubar, this same search can be started from the **Similar Protein Search** box in the Protein page (*vide infra*)
  - **SMILES/SMARTS** searches (see Ligand SMILES or SMARTS Searches Section 5, page 66). Results from these types of searches can be selected to be saved ***before*** or ***after*** the search has been run.

- **2D/3D** searches (see 2D/3D Ligand Substructure Searches Section 6, page 70). Results from this type of search can be selected to be saved ***before*** or ***after*** the search has been run.
- **Similar ligand** searches (see Similar Ligand Searches Section 7, page 78). Results from this type of search can only be selected to be saved ***after*** the search has been run.
- **Similar protein chain** searches (see Similar Protein Chain Searches Section 8, page 81). Results from this type of search can only be selected to be saved ***before*** the search has been run.
- **Binding site superpositions** (see Similar Binding Site Searches (and Superposition) Section 9, page 84). Results from this type of search can only be selected to be saved ***after*** the search has been run.

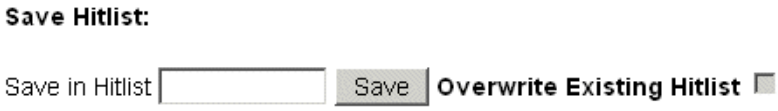
- In order to specify that you would like to save a hitlist before running a search, type the required hitlist name into the *Save in Hitlist* box. The example below shows a protein hitlist (called *ESTERASE*) being saved for keyword search:



- To overwrite a previously saved hitlist, type the hitlist name into the *Save in Hitlist* box, then activate the **Overwrite Existing Hitlist** check box and click **Submit**.
- A hitlist can be saved after a search has been run by clicking on the **Save in Hitlist** button which

will be located in the bottom left frame of the results window (for text, sequence, SMILES and 2D/3D searches) or the top of the ligand similarity results page (for ligand similarity searches).

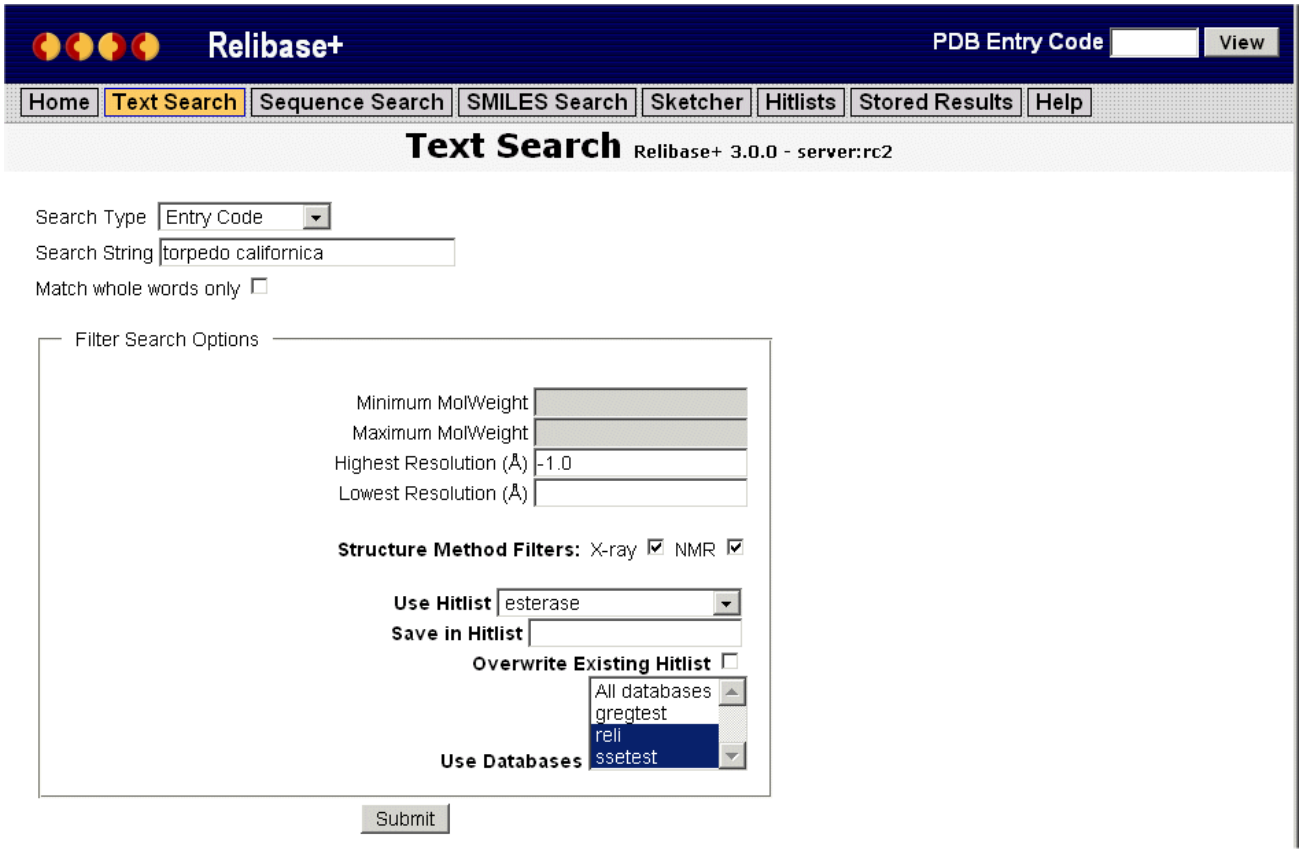
- In the case of similar binding site searches, results can be saved after a search has been run by typing the hitlist name into the *Save in Hitlist* box at the bottom of the similar binding site search results page, then hit **Save**.



- The hitlist type (protein, ligand, or cavity) will be determined automatically according to the search being run.

6.3 Using Hitlists in Subsequent Searches

- The following searches can use hitlists which have been saved from a previous search:
  - **Text** searches (see Keyword Searches Section 3, page 58).
  - **SMILES** searches (see Ligand SMILES or SMARTS Searches Section 5, page 66).
  - **2D/3D** searches (see 2D/3D Ligand Substructure Searches Section 6, page 70).
  - **Sequence** searches (see Protein Sequence Searches Section 4, page 65).
  - **Binding Site Superposition** searches (see Similar Binding Site Searches (and Superposition) Section 9, page 84).
- To use a previously saved hitlist, select the required hitlist from the drop-down menu next to *Use Hitlist*. Only entries in the selected hitlist will be considered in the new search.
- The example below shows a keyword search for entries in the protein hitlist *ESTERASE*:



6.4 Viewing and Editing Hitlists

- Click on the **Hitlists** button on the Relibase+ menubar. This loads three frames into the browser, below the menubar.
- The top-left frame lists the hitlists you have stored according to *Set Name*, *Owner*, *Type* (ligand, protein or cavity), *Size* (number of entries in hitlist), and *Access* state. *Last modification* date and time are also provided:

Set Name	Owner	Type	Size	Access	Remove	Last Modification	Content
berglig	toby_iw_deb4	Ligand	6	Public	Not owner	Mon Aug 11 2008 15:06:28	ASCII XML
tmp_1u4d	adf_rb3	Protein	3	Public	Not owner	Tue Aug 12 2008 13:51:47	ASCII XML
tmp_cav_1ud4	adf_rb3	Cavity	14	Public	Not owner	Wed Aug 13 2008 15:06:55	ASCII XML
esterase	henderson	Protein	683	Private Change	Delete	Mon Aug 18 2008 15:02:12	ASCII XML
hydrolase	henderson	Protein	10591	Private Change	Delete	Mon Aug 18 2008 15:50:20	ASCII XML

- The *Access* state indicates whether a hitlist has a *Private* or *Public* function. *Public* hitlists can be viewed but not edited by other users. *Private* hitlists can be neither viewed nor edited by others. Only list *Owners* can remove/delete lists.
- To view the contents of a hitlist as an *ASCII* or *XML* file, click on the appropriate link under *Content*. The hitlist will be displayed in the format specified in a separate browser window.

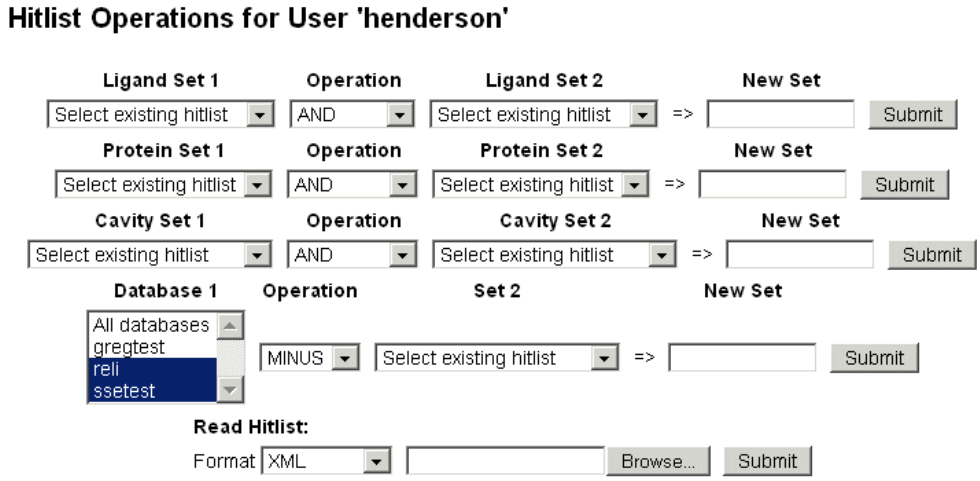
Note: any saved XML hitlist can be read back into Relibase+ (see Loading XML Format Hitlists Section 6.8, page 55).

- To edit the contents of a hitlist click on the name of the hitlist, e.g. *hydrolase*. The hitlist entries will be displayed in the right-hand frame:

- The above example shows protein hitlist entries, for ligand hitlists the 2D ligand chemical diagrams will also be displayed. Output options are also available for ligand hitlists (see Saving Hitlists Section 6.6, page 54).
- Click on the **View Entries** button to re-load the entire hitlist, or select a hitlist entry to link to the corresponding Relibase+ protein entry or ligand page (depending on the hitlist type).
- Use the check boxes to select hitlist entries. Selected entries can be:
  - Added to a different hitlist: select the target hitlist from the popup menu next to *Add to Set* and hit the **Submit** button.
  - Removed from the current or from another hitlist: select the target hitlist from the popup menu next to *Remove from Set* and hit **Submit**.
  - Make into a new hitlist: enter the name of the new hitlist into the text window next to *Make New Set* and hit **Submit**.

6.5 Combining and Translating Hitlists

- Click on the **hitlists** button on the Relibase+ menubar. This loads three frames into the browser, below the menubar. In the bottom-left frame, you can combine or convert different hitlists using logical operators in order to generate new hitlists.



- To combine hitlists using simple logical operators:
  - Select the two hitlists you wish to combine using the pull down menus below the appropriate hitlist type, e.g. *Ligand Set 1* and *Ligand Set 2*. The hitlists to be combined must be of the same type, i.e. either both Ligand, both Protein, or both Cavity.
  - Select the logical operator to be applied on the two hitlists:
    AND means each entry in the new hitlist must occur in both hitlists.
    OR means each entry in the new hitlist must occur in at least one of the two hitlists.
    MINUS means each entry must occur in the first hitlist, but must not occur in the second.
  - Enter the name of the new hitlist in the appropriate *New Set* text box and hit the **Submit** button to create the new hitlist.
- Hitlists of entries can be converted from one type (ligand, protein, or cavity) to another. Convert hitlists as follows:
  - Select the hitlist you wish to convert using the popup menu below either *Ligand Set 1, Protein Set 1*, or *Cavity Set 1*.
  - Select => *Ligand*, => *PDB*, or => *Cavity* where appropriate from the popup menu of logical operators (note that the menu options available reflect the hitlist type selected).
  - Enter the name of the new ligand, protein, or cavity hitlist in the *New Set* text box and hit the **Submit** button to create the new hitlist.
- Hitlists can also be copied or renamed:
  - Select the hitlist you wish to copy or renamed using the pulldown menu below either *Ligand*

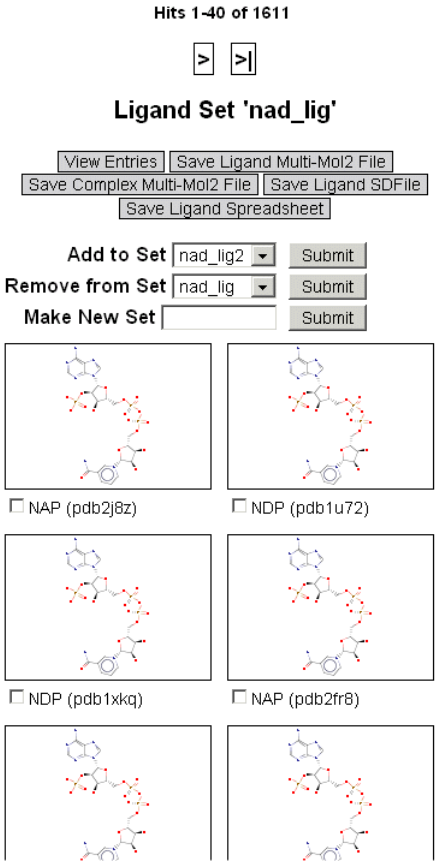
Set 1, Protein Set 1, or *Cavity Set 1*.

- Select the *Copy* or *Rename* option from the corresponding *Operation* pulldown menu .
- Type in the new hitlist name for the copied or renamed list and **Submit**.

- Hitlists may be subtracted from any loaded database:
  - Select the database you wish to subtract the hitlist from the *Database 1* pulldown list.
  - *Minus* will already be selected in the *Operation* pulldown menu.
  - Select the hitlist you wish to subtract from the *Set 2* pulldown menu.
  - Type the name of the new hitlist and press **Submit**. A new hitlist of the same type (i.e. *Protein*, *Ligand* or *Cavity*) will be added to the list.

6.6 Saving Hitlists

- Entries in ligand hitlists can be saved by viewing the hitlist then selecting one of the following options:



- **View Entries**: this button re-loads the search results into the browser window.
- **Save Ligand Multi-Mol2 File**: use this button to save all hit ligands to a multi-mol2 file.
- **Save Complex Multi-Mol2 File**: use this button to save the ligand and its binding site to a multi-mol2 file.

- **Save Ligand SDFile**: use this button to save all ligands to an SDFile.
- **Save Ligand Spreadsheet**: use this button to save ligand information for all hit ligands to a *.csv* (comma separated value) file. Information content of this file includes: PDB code, ligand compound name, number of heavy atoms, empirical formula and the ligand SMILES code.

- If you have more than one hitlist of a given type (e.g. *Protein*, *Ligand* or *Cavity*) these can be added or removed from other sets of the same type and saved (see Viewing and Editing Hitlists Section 6.4, page 51).

6.7 Deleting Hitlists

- Click on the **Hitlists** button on the Relibase+ menubar. This loads three frames into the browser, below the menubar. The top-left frame lists the hitlists you have stored (see Viewing and Editing Hitlists Section 6.4, page 51).
- Click on **Delete** in the column labelled *Remove* to remove a hitlist.

6.8 Loading XML Format Hitlists

- It is possible to read in any XML format hitlist saved out from a previous search.
- Click on the **Hitlists** button on the Relibase+ menubar. This loads three frames into the browser, below the menubar. In the bottom-left frame, there is also a *Read Hitlist* option (see Combining and Translating Hitlists Section 6.5, page 53).
- To read in an XML list, select XML from the *Format* pulldown, click on **the Browse** button next to this, select the appropriate file using the file browser, and click **Submit** to load the hitlist.

6.9 Loading PDB Format Hitlists

- It is possible to read and save a PDB plain text listfile as a new hitlist.
- Click on the **Hitlists** button on the Relibase+ menubar. This loads three frames into the browser, below the menubar. In the bottom-left frame, there is also a *Read Hitlist* option (see Combining and Translating Hitlists Section 6.5, page 53).
- To read in a plain text PDB list, select *PDB code* from the *Format* pulldown, click on **the Browse** button next to this, select the appropriate file using the file browser, and click **Submit** to load the hitlist

CHAPTER 3: RUNNING RELIBASE+ SEARCHES

1 **Browsing Database Entries** (see page 57)
2 **PDB Entry Code Searches** (see page 57)
3 **Keyword Searches** (see page 58)
4 **Protein Sequence Searches** (see page 65)
5 **Ligand SMILES or SMARTS Searches** (see page 66)
6 **2D/3D Ligand Substructure Searches** (see page 70)
7 **Similar Ligand Searches** (see page 78)
8 **Similar Protein Chain Searches** (see page 81)
9 **Similar Binding Site Searches (and Superposition)** (see page 84)

1 Browsing Database Entries

- Click on the **Text Search** button in the Relibase+ menubar.
- Select *Browse Entries* from the *Search Type* pull down menu. Now select the database (i.e. *reli* or an inhouse database) or hitlist to be browsed, and any required resolution or X-ray/NMR filters to be applied.
- Hit the Submit button to view all database or hitlist entries that satisfy any filters set.

2 PDB Entry Code Searches

2.1 Performing an Entry Code Search using the PDB

- Type the required 4-character text string into the *PDB Entry Code* box at the top-right of any Relibase+ page. PDB entry searches are exact match searches which will match on the given string only, i.e. a search on *1et* will not retrieve *1etr*. Note the text string is not case-specific.

- Hit the **View** button to the right of the *PDB Entry Code* box to start the search.
- The results are presented as a single entry frame and the protein is displayed in a *Hermes* window (if automatic visualiser updates are enabled).

2.2 Performing an Entry Code Search using In House databases

- Click on the **Text Search** button in the Relibase+ menu bar.
- Change the *Search Type* to *Entry Code*.Type the required text string into the *Search String* box. Select the database(s) to be searched in the *Use Databases* box and submit as before. Entry code searches are not exact match searches and will find all matches that contain the search string. Select the databases to be searched in the **Use Databases** box.

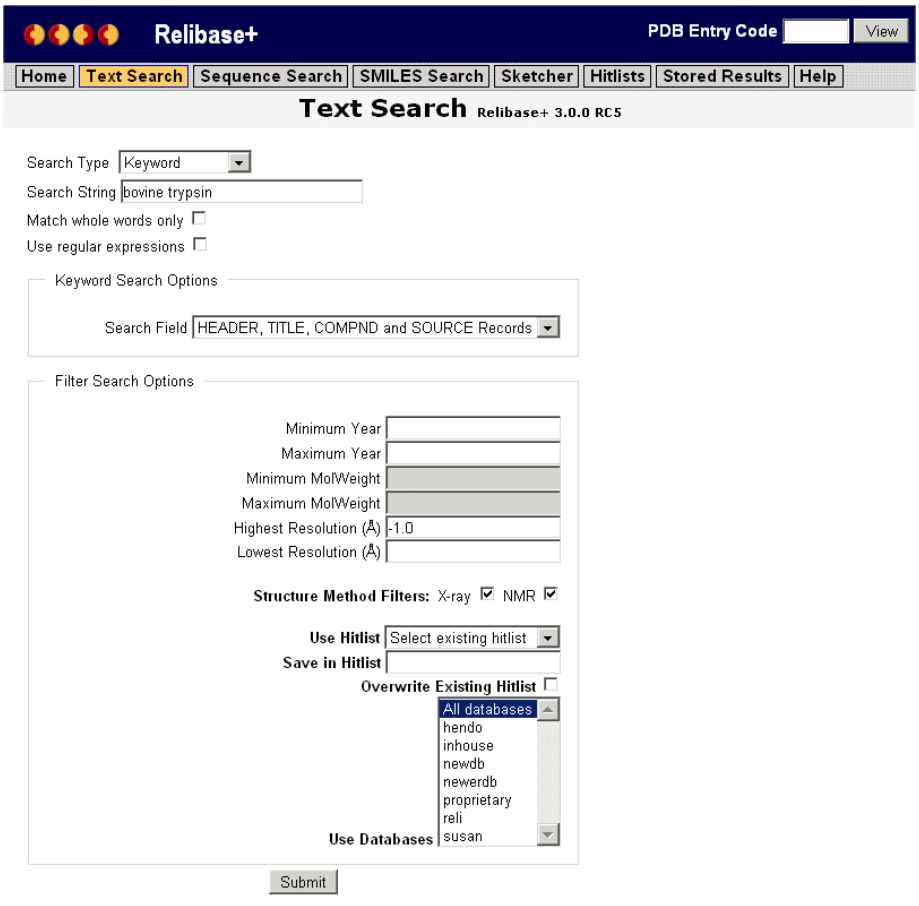
2.3 Hints for Entry Code Searching

- The searches are not case sensitive.
- When using an in-house database, filenames can consist of underscores, alphanumeric characters, hyphens and must start with a letter (either a-z or A-Z). The separation of various parts of the filename for representation in the GUI is in the order of underscores first, then digits and lastly hyphens, for example: *ccdc1mystructAa1* would be represented as *1mystruct1 (ccdc)* and *ccdc_1ets_MQI* would be represented as *1ets_MQI (ccdc)*.
- ***Note***: regular expressions can be used, (see Performing a Keyword Search Section 3.1, page 58).

3 Keyword Searches

3.1 Performing a Keyword Search

- Click on the **Text Search** button in the Relibase+ menubar.
- Select the **Keyword** option in the *Search Type* box.
- By default the *Search Field* will be **HEADER, TITLE, COMPND and SOURCE Records**. A number of options will become available.
- Type the required text string into the *Search String* box.



- Use the **Match whole words only** tick box to gain further control over the results obtained.
- Regular expressions may be used, for example:
  - **"trna guanine transglycosylase":** the use of quotes means that a match will be found for the entire string.
  - ^**trna**: the use of ^ matches the start of the string. In this case, the query would match **trna synthetase** but not **aspartyl trna synthetase**.
  - **.**: matches any character.
  - +: causes the resulting expression to match 1 or more repetitions of the preceding expression. e.g. ab+ will match a followed by any non-zero number of bs; it will not match just a.
  - **[ ]**: square brackets are used to indicate a set of characters. Characters can be listed individually, or a range of characters can be indicated by giving two characters and separating them by a hyphen: **-**. For example, [akm$] will match any of the characters **a**, **k**, **m**, or **$**; **[a-z]** will match any letter.
    Note: regular expression searching is not supported for ligand entry code searches (see Ligand Code Searches Section 3.4, page 62).
- Various additional *Options* are available for all text-based searches. However not all options are available for all searches (see Options for Text-Based Searches Section 3.5, page 64).
- Hit the **Submit** button to start the search.
- The results are presented as a browsable list of Relibase+ entries. The text that makes up the query will be highlighted in red in the *Header*, *Title*, *Compound* and *Source* fields.
- The search results can be saved to a hitlist after the search has been run using the **Save in Hitlist** button in the bottom left-hand frame of the results window.

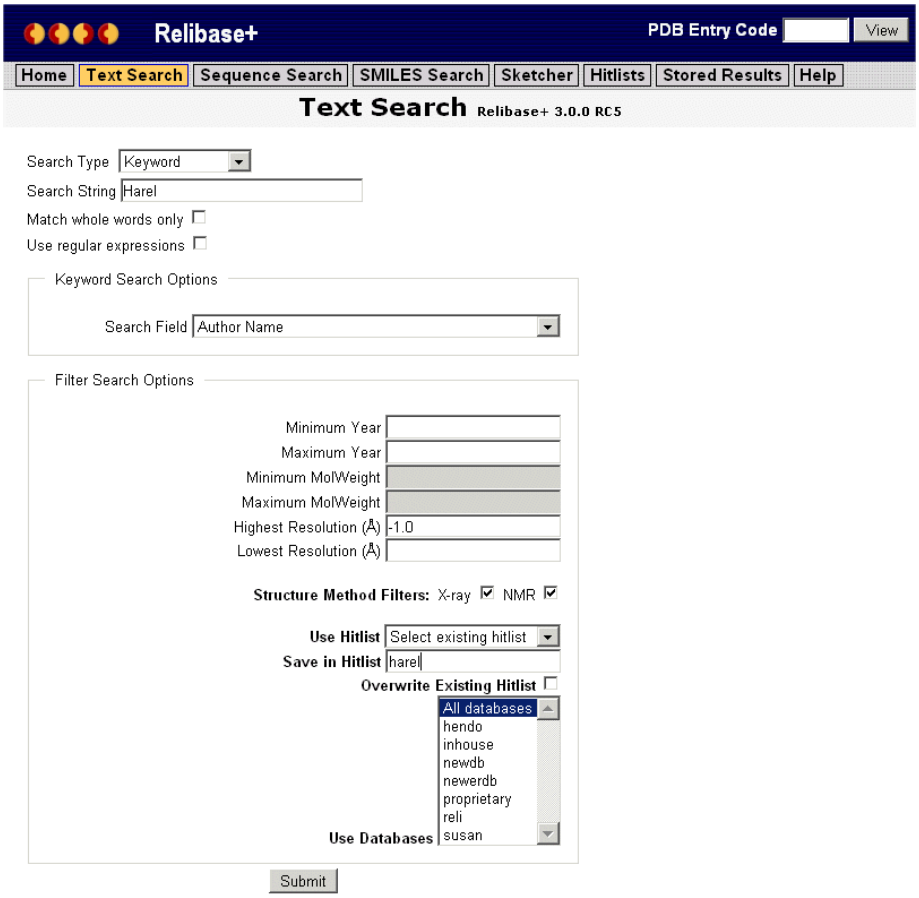
3.1.1 Hints for Keyword Searching in the *Header*, *Title*, *Compound* and *Source* fields

- The searches are not case sensitive.
- The searches are based entirely on the HEADER, TITLE, COMPND and SOURCE records available in the original PDB file. A text-based search using the regular expression "*bovine trypsin"* will not return all bovine trypsin structures, since these two words, separated by exactly one space, are not guaranteed to be present in every PDB bovine trypsin entry. However searching for *bovine trypsin* will bring up hits that contain these two words in any order and position and should contain all the relevant structures, provided that both words (everything in the search string) occur in at least one of the PDB fields noted above.

3.2 Author Name Searches

3.2.1 Performing an Author Search

- Click on the **Text Search** button in the Relibase+ menubar.
- Select the **Keyword** option in the *Search Type* box.
- Select the **Author Name** option from the *Search Field* pull down menu.
- Type the required text string into the *Search String* box.



- Various additional *Options* are available for all text-based searches. However not all options are available for all searches (see Options for Text-Based Searches Section 3.5, page 64).
- Hit the **Submit** button to start the search.
- The results are presented as a browsable list of Relibase+ entries. The author's name that makes up the query will be highlighted in red in the *Author(s)* field.
- The search results can be saved to a hitlist after the search has been run using the **Save in Hitlist** button in the bottom left-hand frame of the results window.

3.2.2 Hints for Author Searching

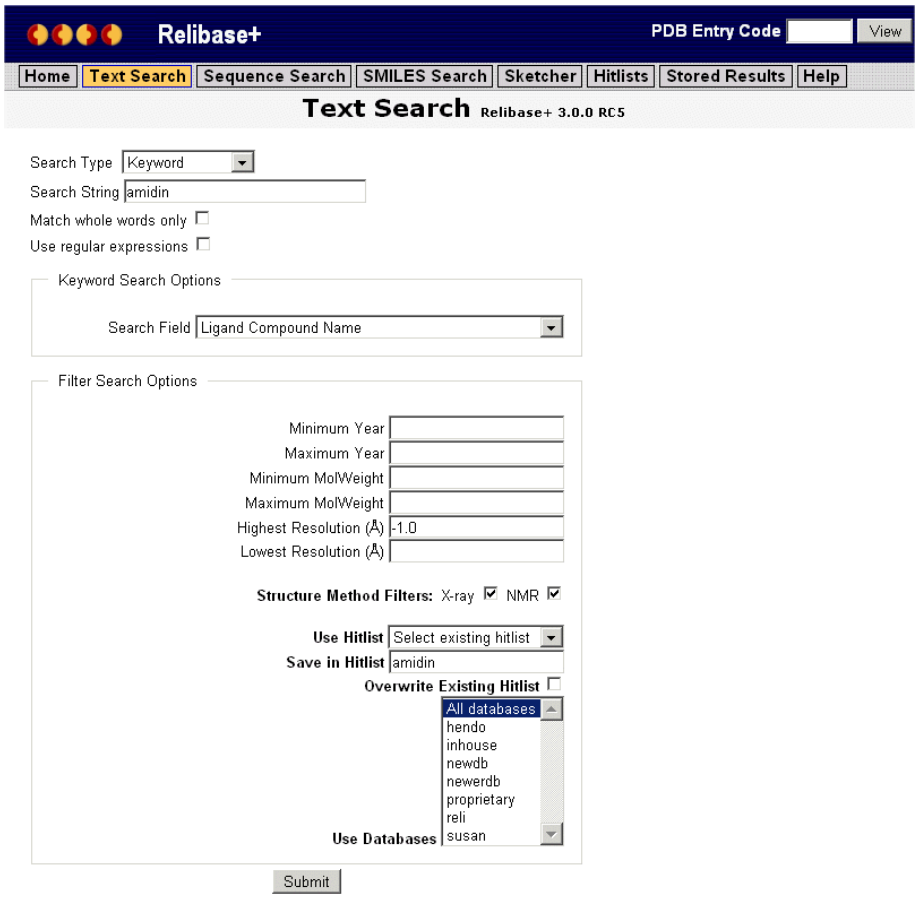
- The searches are not case sensitive.
- Searches for *Huber* will also hit *Glockshuber* unless the *Match whole words only* box is ticked.

- You should be aware that the searches are based entirely on the bibliographic information given in the original PDB file.
- If you wish to include authors' initials, you should use *M.Harel*, i.e. with no space between the initial(s) and the surname.
- It is possible to search for two authors's names simultaneously. Multiple author names should be separated by a space.

3.3 Ligand Compound Name Searches

3.3.1 Performing a Ligand Name Search

- Click on the **Text Search** button in the Relibase+ menubar.
- Select the **Keyword** option in the *Search Type* box.
- Select the **Ligand Compound Name** option from the *Search Field* pull down menu.
- Type the required text string into the *Search String* box.



- Various additional *Options* are available for all text-based searches. However not all options are available for all searches (see Options for Text-Based Searches Section 3.5, page 64).

- Hit the **Submit** button to start the search.
- The results are presented as a browsable list of ligands. The search text will be highlighted in red within the *Chemical name* of each ligand. The ligand and corresponding binding site are displayed in the *AstexViewer* window.
- The search results can be saved to a hitlist after the search has been run using the **Save in Hitlist** button in the bottom left-hand frame of the results window.

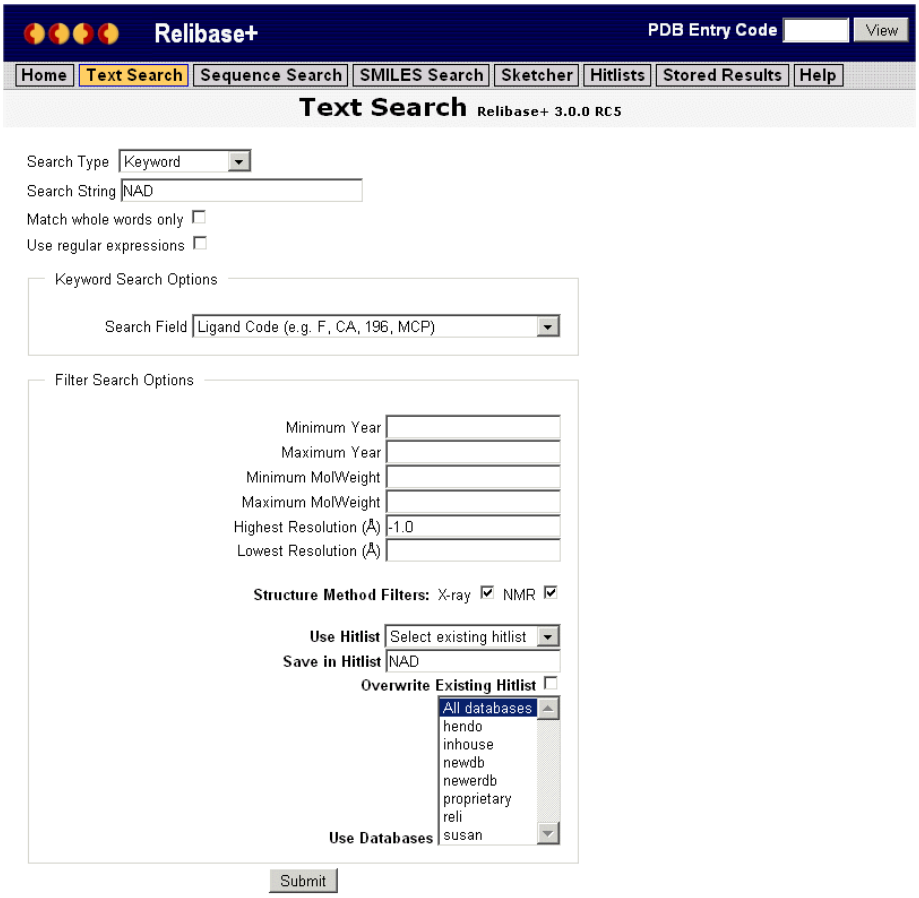
3.3.2 Hints for Ligand Name Searching

- The searches are not case sensitive.
- Ligand name searching can be useful as a quick way of finding examples of a particular type of structure, since it is often quicker to type a name than draw a substructure. However, substructure searching is usually better if you want to be sure of finding all examples of a particular type of ligand, since ligands may be named in different ways.
- In general, and particularly in locating natural products, search for only the key *root* part of the name, e.g. *picolin, penicill*. This is because the names may have derivative endings.
- Searches for trivial names, drug names etc. can be useful. The trivial name is usually the only name stored for natural products.

3.4 Ligand Code Searches

3.4.1 Performing a Ligand Code Search

- Click on the **Text Search** button in the Relibase+ menubar.
- Select the **Ligand Code** option from the *Search Field* pull down menu.
- Type the required text string into the *Search String* box.



- One, two and three letter ligand entry codes can be searched using this method, e.g. *Ca* will return all PDB entries with calcium metal ions present, and *F* will return all PDB entries with fluorine counter-ions. Note that all characters in the search string will be matched (i.e. this is not a substring match; *AC* will not match *ACE*, but regular expressions can be used e.g. *AC.*).
- Various additional *Options* are available for all text-based searches. However not all options are available for all searches (see Options for Text-Based Searches Section 3.5, page 64).
- Hit the **Submit** button to start the search.
- The results are presented as a browsable list of ligands. The search text will be highlighted in red within the *Chemical name* of each ligand. The ligand and corresponding binding site are displayed in the *AstexViewer* window.
- The search results can be saved to a hitlist after the search has been run using the **Save in Hitlist** button in the bottom left-hand frame of the results window.

3.4.2 Hints for Ligand Entry Code Searching

- The entry codes are assigned to so-called HET groups in the structure. A ligand can be built up of more than one HET group (each with its own entry code). One common situation is where the ligand is a polypeptide chain, in which case each amino-acid in the chain is a HET group and is

represented by the standard 3-letter code for amino-acids.

- The searches are not case sensitive.
- There are ambiguities in some entry codes of HET groups. MAL, for example is used for malonate anions, but also for maltose.
- As with ligand name searches, it is often safer to use a substructure search instead.

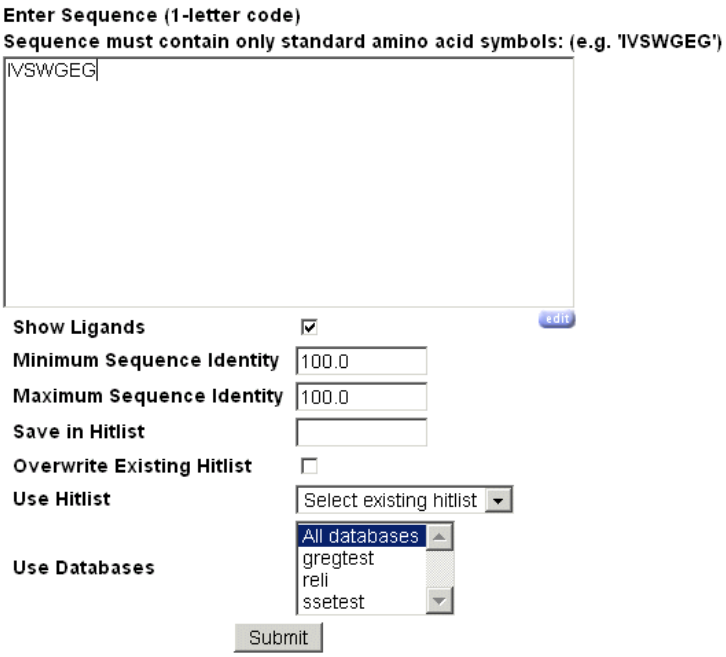
3.5 Options for Text-Based Searches

- Various additional *Options* are available for all text-based searches. However not all options are available for all searches:
  - *Minimum Year* and *Maximum Year* boxes can be used to restrict searches based on publication year. Leaving the boxes empty means that all years will be considered. To return hits from only one year, e.g. 1995, enter 1995 into both *Minimum Year* and *Maximum Year* boxes. These options are available for all searches.
  - *Minimum MolWeight* and *Maximum MolWeight* boxes can be used to restrict the molecular weight (and size) of ligands that you wish to retrieve from your search. Leaving these boxes empty means that all ligands are considered. These options are only available if a *Ligand Compound Name* or a *Ligand PDB Code* search is carried out.
  - *Highest Resolution* and *Lowest Resolution* boxes can be used to filter on the experimental precision of X-Ray derived structures that you wish to consider. If you only wish to consider X-Ray structures with a resolution of 2.0Å or better then enter *2.0* into the *Lowest Resolution* box. If only the X-Ray Structure Method Filter is toggled on, then a *Highest Resolution* can also be set. All NMR structures by default have a resolution set of -1.0, and cannot be filtered in this way.
  - *Structure Method Filters* can be used to restrict the search to either X-Ray or NMR derived structures. The default is that no restriction is made. Toggle off the criterion that is not required.
  - *Use Hitlist* allows you to to use a previously identified list which can be selected from the pop-up menu next to *Use Hitlist*. The default is *Select existing hitlist*; until a hitlist is selected, the entire set of database(s) will be searched.
  - *Save in Hitlist* allows you to save the results of a search in a hitlist; type the required hitlist name into the *Save in Hitlist* box before you start the search. You will not be allowed to overwrite an old hitlist unless you click on the *Overwrite Existing Hitlist* button.
  - *Use Databases* allows you to select which database or combination of databases is searched. The default setting is to search *All databases*.

4 Protein Sequence Searches

4.1 Performing a Sequence Search

- Click on the **Sequence Search** button in the Relibase+ menubar.
- Type the required one-letter-code amino acid sequence into the *Sequence Search* box.



- Various *Options* are available:
  - If you wish to display the ligand 2D chemical diagrams in the resulting list of chains, select the **Show Ligands** check box (this is the default).
  - *Minimum Sequence Identity* and *Maximum Sequence Identity* boxes can be used to specify the required sequence identity as a percentage with respect to the reference chain (default is 100%).
  - *Save in Hitlist* allows you to save the results of a search in a hitlist; type the required hitlist name into the *Save in Hitlist* box before you start the search. You will not be allowed to overwrite an old hitlist unless you click on the *Overwrite Existing Hitlist* button. Note that it is not possible to save the search results to a hitlist after the search has been run.
  - *Use Hitlist* allows you to to speed the search by restricting the data covered to a previously identified list which can be selected from the pop-up menu next to *Use Hitlist*. The default is *Select existing hitlist*; until a hitlist is selected, the entire set of databases will be searched.
  - Similarly *Use Databases* allows you to select which database or combination of databases is searched. The default setting is to search *All databases*.
- Hit the **Submit** button to start the search.

- The results are displayed as a list of chains. The ordering of the list is set using the Smith-Waterman score. This takes account of the sequence identity, the number of residues this sequence identity applies to, and, in cases where a match is impossible without the inclusion of insertions in the matched sequence, the number of insertions that are required. Chains with maximum identity, number of homologous amino acids, and fewest insertions, are ranked highest.
- The search results can be stored by typing a name and description of the search into the relevant boxes in the *Save Similar Sequence Results* section of the results window (at the bottom of the page), then hitting **Save**. The stored search can be viewed at a later date via the *Stored Results* window (see Creating Hitlists Section 6.2, page 48).
  Note: the Fasta sequence search finds the 1000 best sequence matches in the sequence files, then filters the resulting chains on homology. When searching for low homologies, the required chains may not be within the 1000 best sequence matches found by Fasta.

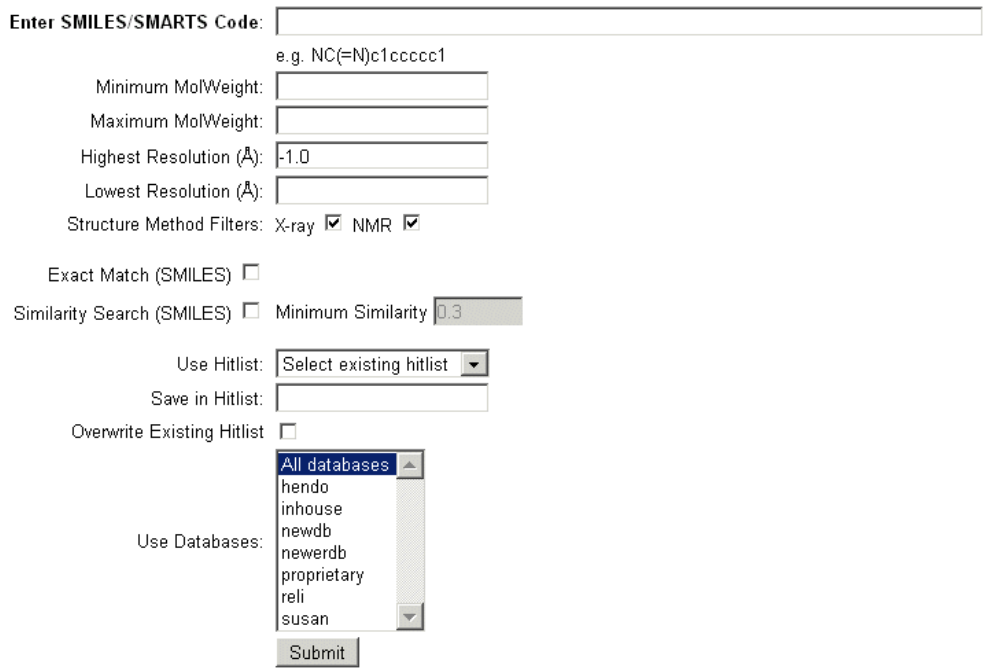
4.2 Hints for Sequence Searching

- The searches are not case sensitive.
- The searches are done via the *FastA* program. Only those hits are returned that are considered significant by *FastA*. If given long strings *FastA* will also return hits for which only a subset of the original search string is matched. For a detailed description of *FastA*, the user is referred to the *FastA* user manual (see *http://fasta.bioch.virginia.edu/*).
- Information on installing FastA is provided in the Relibase+ installation notes (*http://www.ccdc.cam.ac.uk/support/documentation/#relibase*).

5 Ligand SMILES or SMARTS Searches

5.1 Performing a SMILES/SMARTS Substructure Search

- Click on the **SMILES Search** button in the Relibase+ menubar.
- Type the required Smiles or Smarts string of the substructure you wish to search into the *Enter SMILES/SMARTS Code* box.



- Various *Options* are available
  - *Minimum MolWeight* and *Maximum MolWeight* boxes can be used to restrict the molecular weight (and size) of ligands that you wish to retrieve from your search. Leaving these boxes empty means that all ligands are considered.
  - *Highest Resolution* and *Lowest Resolution* boxes can be used to filter on the experimental precision of X-Ray derived structures that you wish to consider. If you set a highest resolution of anything other than -1.0 or “empty”, no NMR derived structures will be retrieved.
  - *Structure Method Filters* can be used to restrict the search to either X-Ray or NMR derived structures. The default is that no restriction is made. Toggle off the criterion that is not required.
  - *Exact Match (SMILES)*: activate this check box when you wish to retrieve ligands containing only the exact query SMILES string.
  - *Similarity Search (SMILES)* (see Performing a SMILES Similarity Search Section 5.3, page 70).
  - *Use Hitlist* allows you to use a previously saved hitlist which can be selected from the pop-up menu next to *Use Hitlist*. The default is *Select existing hitlist*; until a hitlist is selected, the entire database will be searched.
  - *Save in Hitlist* allows you to save the results of a search in a hitlist; type the required hitlist name into the *Save in Hitlist* box before you start the search. You will not be allowed to overwrite an old hitlist unless you click on the *Overwrite Existing Hitlist* button. Sequence search results can also be saved after the search has been run (see Creating Hitlists Section 6.2, page 48).

- *Use Databases* allows you to select which database or combination of databases is searched. The databases must first have been loaded when the Relibase+ server was started (see *Section 3* of the Inhouse Data Processing manual). The default setting is to search *All databases*.
- Hit the **Submit** button to start the search.
- The results are displayed as a list of ligands.

5.2 The Use of SMILES/SMARTS in Relibase+

SMILES are string representations of 2D molecules, while SMARTS are string representations of substructures. SMARTS provide variable atom/bond properties and atom/bond constraints which are not part of SMILES. Detailed information on both can be found on the Daylight web pages (*http://www.daylight.com/dayhtml/doc/theory/index.html*). Guidelines about the use of SMILES and SMARTS are given in the sections that follow:

5.2.1 SMILES searching

The following information is helpful if you use SMILES in Relibase+:

- Information about charges, isotopes and stereochemistry is ignored.
- Hydrogens are only allowed in brackets together with a heavy atom (e.g. [NH3] or [OH]).
- Hydrogens can be used to fill up valencies (e.g. C(=O)[NH2] will find only carbamoyl groups, and not e.g., peptide linkages).
- Relibase+ supports the bond-type 'any' (Symbol: '~').
- Relibase+ supports three types of atom 'wildcards':
  - A: Any atom. This will only match hydrogen if there are hydrogen atoms stored explicitly in the ligand. This is not always the case.
  - R: Any atom except H-Atoms
  - X: Any atom except C- and H-Atoms
- Designation of aromaticity using lower case letters is supported for 5- and 6-membered aromatic rings. Use single and double bonds for others e.g. unstaturated 5-membered rings. The SMILES code ':' can be used to designate a single aromatic bond if necessary.
- Relibase+ does not support tautomeric states. Use bonds of type 'any' (SMILES code ~).
- Queries using '.' are not supported.

5.2.2 SMARTS searching

The implementation of Smarts in Relibase+ is not comprehensive; limitations are primarily due to the way in which ligands are stored in Relibase+. The following should be taken into consideration when using Smarts:

- Relibase+ assumes bond types given in the SMARTS query match Relibase+ conventions. In

particular:

- Six-membered aromatic rings have aromatic bond types, however complete 6-membered rings in SMARTS input with single-double bond types will be converted to aromatic.
- Five-membered rings are non-aromatic unless pi bonded to a metal (e.g. ferrocenes).

- Due to the nature of the data source, hydrogen counts on atoms other than carbon are not reliable, use of *Dn* atom constraint (number of non-hydrogen connections) is recommended rather than *Xn* (total number of connections) for heteroatoms.

Unsupported features (general):

- Dot disconnected fragments, e.g. `(C).(C)`
- Recursive SMARTS, e.g. `[\$(CC);\$(CCC)]`
- Reaction SMARTS, e.g. `[CC>>CC]`.

Unsupported features (atom properties):

- Some atom constraints (where n is an integer):
  - `v<n>`: valency constraint.
  - `x<n>`: number of ring connections constraint.
  - `h<n>`: implicit hydrogen constraint (no distinction is made between implicit and explicit H in Relibase+).
  - Charge constraints (no charges are stored in Relibase+).
  - `R<n>` where n>=1 (no smallest set of smallest rings implementation).
  - `\#n`: atomic number (the element symbol should be used).
  - `<n>`: atomic mass.
- Stereochemical descriptors.
- Constraints of different types combined with OR operator, e.g. `[X1,D2]`.
- High precedence AND in OR subexpression, e.g. `[C,N&H1]` (constraints can only be applied to all element types in an atom).

Unsupported features (bond properties):

- Stereochemical descriptors for double bonds: these are treated as single bonds with unspecified stereochemistry.
- High-precedence AND in OR subexpression, e.g. `=\&\@,-` (cyclic double or single and unspecified cyclicity).
- The following constructs are not supported:
  - NOT any bond, e.g. `!~`.
  - different bond types combined with AND operator, e.g. `-\&=` (single and double).
  - different NOT bond types combined with OR operator, e.g. `!-,!=` (not single or not double,

equivalent to any bond).

5.3 Performing a SMILES Similarity Search

- Click on the **Similarity Search (SMILES)** toggle box. This will activate the *Minimum Similarity* box.
- Choose a minimum similarity threshold between 0 and 1. The default is 0.3. You will find that the longer the SMILES string you wish to match, the higher you will need to set the threshold to avoid returning too many hits. Some trial and error may be necessary. The similarity threshold is a Tanimoto coefficient. Tanimoto coefficients are calculated from the comparison of topological fingerprints of each ligand against that of the reference ligand.The results are displayed as a list of ligands ordered in terms of similarity to the SMILES substructure. The Tanimoto similarity coefficient is given for each ligand.
  Note: It is not recommended to use the bond-type 'any' (Symbol: '~'), or the three atom wildcards, A, R and X, in similarity searches, as the matching of these symbols is not supported for that use.

6 2D/3D Ligand Substructure Searches

6.1 2D Substructure Searching

- A 2D substructure search is most usually carried out to find ligands which include the 2D substructure that is of interest.
- There are four molecule types in Relibase+, **Protein**, **Nucleic Acid, Ligand** and **Water**.
- A 2D substructure search must be used to find proteins with *unusual* amino acids. The definition of a **Protein** in Relibase+ encompasses *normal* (ie. the 20 most commonly occurring AAs) amino acids only. Obviously most protein substructure search that do not involve these *unusual* residues can be more easily carried out via a *Sequence Search* (see Performing a Sequence Search Section 4.1, page 65).
- Nucleic acids cannot be searched via a substructure search.
- A 2D ligand substructure search is often worth doing prior to a 3D substructure search, in order to reduce the size of the search list that will be queried by the 3D search.

6.2 Hints for 2D Substructure Searching

- If you are unfamiliar with chemical substructure searching, try a few simple searches, e.g. for 6-membered carbocyclic rings, 4-coordinate transition metals.
- Finding exact structures requires complete definition of the target molecule, including H-atoms. If the Relibase+ database does not contain the target, then relax the H-atom specification to see if simple derivatives are present.
- In an initial search, do not over-specify the substructure, e.g. in terms of allowed substitution. It

is better to get too many hits and then impose tighter chemical constraints. Let the database tell you what it contains!

- If you are unsure of the bond type used in the Relibase+ database for a particular substructure, use the bond type *any* and look at the resulting hits in order to formulate a more precise query.

6.3 3D Substructure Searching

- A 3D substructure search is one in which geometric constraints are added to a 2D substructure search so that only certain geometries are represented in the hits retrieved.
- The geometric restraints may be either **Distance, Angle** or **Torsion Angle** restraints.
- Geometric parameters may be defined using just atoms or atoms and objects (see Geometric Parameters Section 11, page 139).

6.3.1 3D Ligand Substructure Searching: Constraining geometry

- There are times when you wish to search for ligands which only conform to certain geometries. For instance you may wish the two ends of the ligand to lie within a certain distance, or you may wish to only find ligands which exhibit a certain type of intramolecular hydrogen bond.

6.3.2 3D Ligand Substructure Searching: Monitoring geometry

- Another use for 3D ligand substructure searching is to monitor certain geometrical parameters. This is useful for e.g.:
  - Generating geometry histograms.
  - Locating substructures with specific conformations (torsion angles).
  - Locating specific metal coordination geometries, e.g. tetrahedral rather than square planar.
- Geometry histograms can be accessed once a search is completed. When a 3D substructure search is completed, a **histogram(s)** link appears at the bottom left hand corner of the *Substructure Search Result* page. Clicking on this link brings up histograms of frequency against geometry, for the geometry parameters defined in the search (see Viewing Distribution Histograms for Geometrical Parameters Section 3.5.1, page 25).

6.4 Basic Guide to 3D Ligand Substructure Searching

- Draw the required substructure **(see CHAPTER 5: USING THE RELIBASE SKETCHER**, page 119)**.**
- If necessary, define geometric objects such as centroids (see Defining Geometric Objects Section 10.2, page 138).
- Define the required geometric parameters (i.e. the parameters you want to constrain). These may involve just atoms or atoms and objects (see Geometric Parameters Section 11, page 139).
- When you define each parameter, constrain it as required, e.g. specify a range of distances for a

bond length (see Applying Constraints Section 12, page 144).
Note: if you wish to monitor a parameter rather than constrain it, then you will normally have to edit the constraint in order to enlarge the allowed range to encompass that you wish to monitor. The maximum distance in a distance constraint, for instance, is set at a default of only 3.5Å.

- Run the search (see Running a Search Section 6.8, page 74).

6.5 Hints for 3D Substructure Searching and Tabulating Geometries

- To learn how to run searches involving protein and/or water molecules, read the section on nonbonded interactions (see Non-bonded interaction searching Section 6.6, page 72).
- Geometric parameters are often defined so that they can be analysed later, using histograms.
- All required geometric parameters must be defined in the drawing window when the substructure is drawn. They cannot be defined after a search has been run.
- Think carefully about the problem being studied to ensure that you have specified geometric parameters that adequately describe that problem. The obvious choice is sometimes not the best.
- Once defined, any geometric parameter can be used as a search constraint by specifying suitable limiting values.
- In setting geometric constraints, it is often useful to survey typical values found in the Relibase+ database before deciding the limiting values to be used in a subsequent search.
- If you have drawn a complicated query, e.g. with multiple distance constraints, the search may be slow. To check whether the hits you retrieve are of the desired type, you can interrupt the search after it has found a few hits. This enables you to inspect the hits found so far.

6.6 Non-bonded interaction searching

- A highly useful facility is the ability to search on non-bonded interactions between a ligand and a protein or between two proteins. Such searches can be set up also involve water molecules. These kind of searches are useful for e.g.:
  - Finding particular types of interactions between proteins and ligands, e.g. hydrogen bonds, contacts to metals, etc.
  - Generating tables of nonbonded interaction geometries.
  - Finding a particular arrangement of amino acid residues, e.g. catalytic triad (SER-HIS-ASP).
  - Looking for water mediated ligand-protein interactions.

6.6.1 Basic Guide to Nonbonded Protein-Ligand Interaction Searching

- Draw the required substructures (see CHAPTER 5: USING THE RELIBASE SKETCHER, page 119).
- Ensure that the *MoleculeType* is set correctly for each substructure (see Setting Molecule Types Section 1.3, page 120).
- Make sure at least one of the substructures is of *MoleculeType* **Ligand**.
- Make sure each substructure is used in the definition of at least one distance constraint.
- Run the search (see Running a Search Section 6.8, page 74).

6.6.2 Basic Guide to Nonbonded Protein-Protein Interaction Searching

- Draw the required substructures (see CHAPTER 5: USING THE RELIBASE SKETCHER, page 119).
- Ensure that the *MoleculeType* is set to **Protein** for each substructure (see Setting Molecule Types Section 1.3, page 120).
- Make sure each substructure is used in the definition of at least one distance constraint (see Applying Constraints Section 12, page 144).
- Run the search. (see Running a Search Section 6.8, page 74).

6.6.3 Hints for Nonbonded Interaction Searching

- When searching for nonbonded protein-ligand interactions at least one of the substructures must be of the molecule type **Ligand**. At least one of the distance constraints must involve a **Ligand** atom.
- When searching for nonbonded protein-protein interactions ensure that the molecule type is set to **Protein** for each substructure.
- All atoms in a given substructure (i.e. a part of the query linked by covalent bonds) must be of the same Molecule type.
- Water can be included in any nonbonded interaction search.
- Multiple distance constraints are treated as logical *AND* operators. If you define two distance constraints, e.g. between a **Ligand** substructure and a **Protein** substructure, both have to be fulfilled simultaneously.
- All required geometric parameters must be defined in the drawing window when the substructure is drawn. They cannot be defined after a search has been run.
- If you have drawn a complicated query, e.g. with multiple distance constraints, the search may be slow. To check whether the hits you retrieve are of interest, you can interrupt the search after it has found a few hits and to inspect the hits found thus far.
- The order the contact atoms are drawn in is important; you may obtain different search results depending on which atom is drawn first. For example, the search for a contact between an Fe

atom and ligand O atoms.

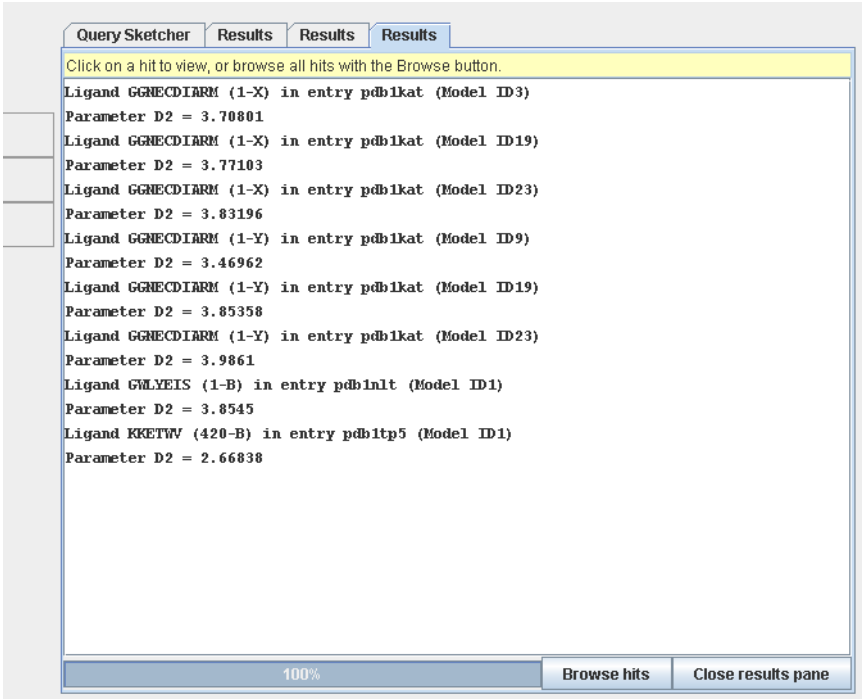
- In cases where the O is drawn first: if the search returns two O groups in the same ligand that coordinate the Fe, only one hit would be added to the hitlist (i.e. only the first ligand hit); if the search returns 2 different ligands that contain O atoms coordinating the Fe, two hits would be added to the hitlist (i.e. one hitlist entry for each ligand).
- If the Fe atom is drawn first and the search returns two different ligands that contain an O atom coordinating the Fe, one hit is returned in the hitlist (i.e. only one contact to a given Fe atom is added to the hitlist).

6.7 Drawing a 2D/3D Substructure

- For instructions on substructure drawing, refer to the section on using the Relibase+ sketcher (see CHAPTER 5: USING THE RELIBASE SKETCHER, page 119).
- For information on setting up 3d constraints please refer to the relevant section in the chapter on using the Relibase+ Sketcher (see Applying Constraints Section 12, page 144).

6.8 Running a Search

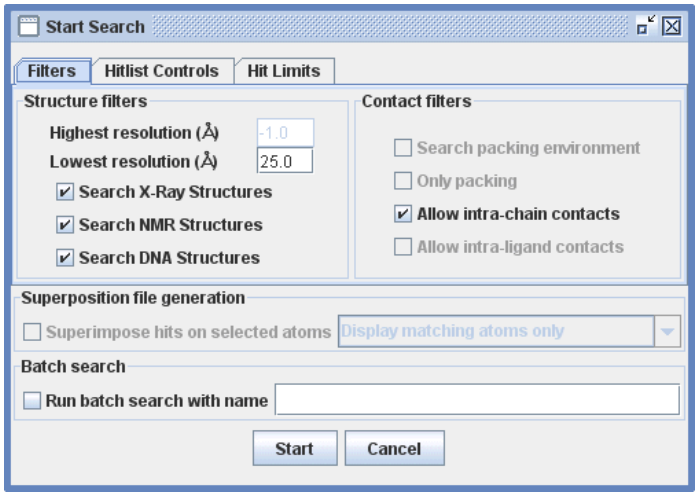
- Having created a 2D or 3D substructure query in the Sketcher, click on the **search** button on the left hand side of the Sketcher window.
- The *Start search* dialogue box will come up. Hitting the **Start** button initiates the search. As the search progresses hits are displayed in a new Sketcher *Results* pane. In addition any 3D parameters are also displayed. Clicking on any line in the *Results* pane links through to the appropriate Ligand page.


Query Sketcher
Results
Results
Results
Click on a hit to view, or browse all hits with the Browse button.
Ligand GGNECDIARM (1-X) in entry pdb1kat (Model ID3)
Parameter D2 = 3.70801
Ligand GGNECDIARM (1-X) in entry pdb1kat (Model ID19)
Parameter D2 = 3.77103
Ligand GGNECDIARM (1-X) in entry pdb1kat (Model ID23)
Parameter D2 = 3.83196
Ligand GGNECDIARM (1-Y) in entry pdb1kat (Model ID9)
Parameter D2 = 3.46962
Ligand GGNECDIARM (1-Y) in entry pdb1kat (Model ID19)
Parameter D2 = 3.85358
Ligand GGNECDIARM (1-Y) in entry pdb1kat (Model ID23)
Parameter D2 = 3.9861
Ligand GWLYEIS (1-B) in entry pdb1nlt (Model ID1)
Parameter D2 = 3.8545
Ligand KKETWV (420-B) in entry pdb1tp5 (Model ID1)
Parameter D2 = 2.66838
100%
Browse hits
Close results pane

- On completion of the search the protein entry browser (see Using the Protein Entry Browser Section 3.2, page 20), or ligand browser (see Using the Ligand Browser Section 3.3, page 22) will open automatically. From here individual hits can be selected for viewing.

6.8.1 Options available on Search: Filters

- There are several options are available in the *Start search* dialogue box. These can be found on either on the *Filters* or the *Hitlist Controls* tabbed pages. It is the *Filters* tabbed page that is displayed by default.



- *Highest Resolution* and *Lowest Resolution* boxes can be used to filter on the experimental precision of X-Ray, DNA or NMR-derived structures that you wish to consider. If you only wish to consider X-Ray structures with a resolution of 2.0Å or better then enter *2.0* into the *Highest Resolution* box. If only *Search X-Ray Structures* is toggled on, then a *Lowest Resolution* can also be set. All NMR structures, by default, have a resolution set of -1.0, and cannot be filtered in this way.
  Note: when restricting a search to DNA structures, either the *Search X-Ray Structures* or *Search NMR Structures* tick box must be activated.
- *Structure Method Filters* can be used to restrict the search to either X-Ray, DNA or NMR-derived structures. The default is that no restriction is made. Toggle off the criterion that is not required.
- If you have created a search which contains two fragments, either ligand-ligand or ligand-protein, then the **Search packing environment** check box in the *contact filters* area will become active. If this box is checked then it means that the search will also include the situations where one of the fragments is part of a neighbouring protein packed closely with the binding site.
  Note: one of the fragments has to be a ligand. Also it will be necessary for a 3D constraint to

be set up between the two fragments, in order for Relibase+ to carry out the search. This constraint can be set to be loose however if needed.

- If the **Search packing environment** box is checked then the **Only packing** check box becomes activated. Checking this box means that the search will only consider cases where the second protein or ligand fragment is part of a neighbouring protein packed closely with the binding site.
- If the query contains two protein fragments then the **Allow intra-chain contacts** check box becomes activated. By default this check box is toggled on. If however it is desired to search only for contacts between two protein chains separately identified in the pdb entry, then this box should be toggled off.
- If the query contains two ligand fragments then the **Allow intra-ligand contacts** check box becomes activated. The default behaviour is for contacts to be found only between two different ligands within the same pdb entry. Toggle this box if you wish also to consider intra-ligand contacts
- If you want to create an overlay of the hits, you must select at least three or more non-hydrogen atoms you wish to superimpose before starting the search. Then, in the *superposition file generation* section of the *Start Search* dialogue box, toggle on the **Superimpose hits on selected atoms** button. You will be given a choice of three options in the pull-down menu to the right. These are, respectively:

  Display matching atoms only. Use this option to superimpose only those substructure atoms drawn in the Sketcher query page.

  Display matching chains. Use this option to display the complete ligand structures after superimposition. Note that complete residues for any protein atoms in the query are shown as well as complete ligands.

  Display entire binding site. This option clearly displays all the atoms, ligand and others, within 6Å of each superimposed ligand. After choosing the appropriate option above, the search is run by hitting Start. Once the search completes the superposition is loaded into the embedded visualiser (AstexViewer) or can be loaded into Hermes via the **Show in Hermes** button. The superposition can be re-loaded using the *Hit Superposition* hyperlink in the bottom left frame. ***Note***: using the *Display entire binding site* option can very rapidly lead to an unintelligible display as the number of superimposed ligands increases. The complexity of the display is also affected by the degree of symmetry of the query. Note also that this option is only available for queries containing ligands.

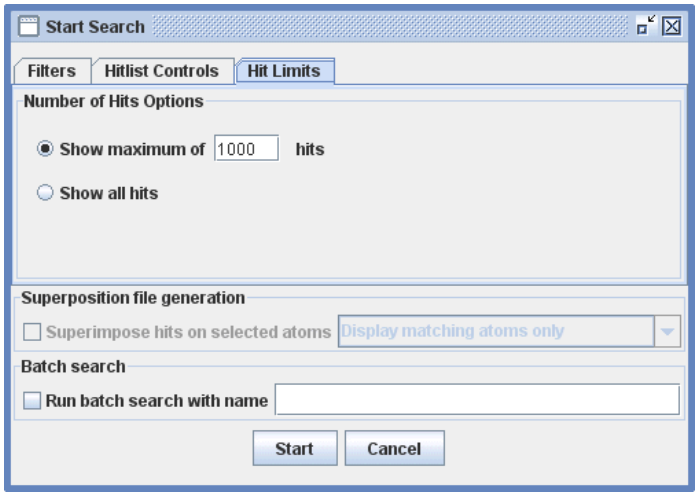
- If you would like to run a search in the background, without the sketcher *Results* tab being updated, activate the *Run batch search with name* tickbox, enter a search name then start the search. Batch search results can be viewed as the search is running using the **Browse Hits**

button in the Results tab, or when the search is finished by loading the search via the *Stored Search Results* pulldown menu in the *Stored Results* window (see Storing Search Results Section 6.1, page 47).

6.8.2 Options available on Search: Hitlist controls

- *Restrict search to hitlist named* allows you to use a previously saved hitlist. Type the name of the required hitlist into the *Restrict search to hitlist named* box in order to restrict the search to ligands or PDB entries in that hitlist.
- *Save search in hitlist named* allows you to save the results of a search in a hitlist; type the required hitlist name into the *Save search in hitlist named* box before you start the search (see Storing, Combining and Converting Search Results Section 6, page 47).
- In the resulting dialog box, various options are available:
  - If atoms were selected (3 or more non-hydrogen atoms must be selected), the dialog box also contains the **Superimpose Hits on Selected Atoms** check box. Click on this if you wish to generate an overlay of the hits (see Viewing 3D Substructure Search Results; Geometrical Analysis Section 3.5, page 25). Selection of this check box generates a pull-down menu from which you can choose to **Display matching atoms only**, **Display matching chains** or **Display entire binding site**. Once the search has been run, the superimposed hits can be viewed on the results page (default) or loaded into Hermes (click on the **Show in Hermes** button).
  - If no atoms were selected, then the only options available are the **Submit** button to start the search or **Cancel** if you want to return to the drawing area.
- The progress of the search is displayed in the *Messages* box below the drawing area. Clicking on any of the hits displayed in the *Hits* box, while the search is progressing, will open a new browser displaying the ligand entry page of the selected hit.
- To interrupt a search, click on the **Interrupt Query** button.
- Hits are loaded in a second browser window and displayed as a browsable list of ligands (see Using the Ligand Browser Section 3.3, page 22). If you interrupted the search, all hits found thus far will be shown.

6.8.3 Options Available on Search: Hit Limits in Substructure Searches



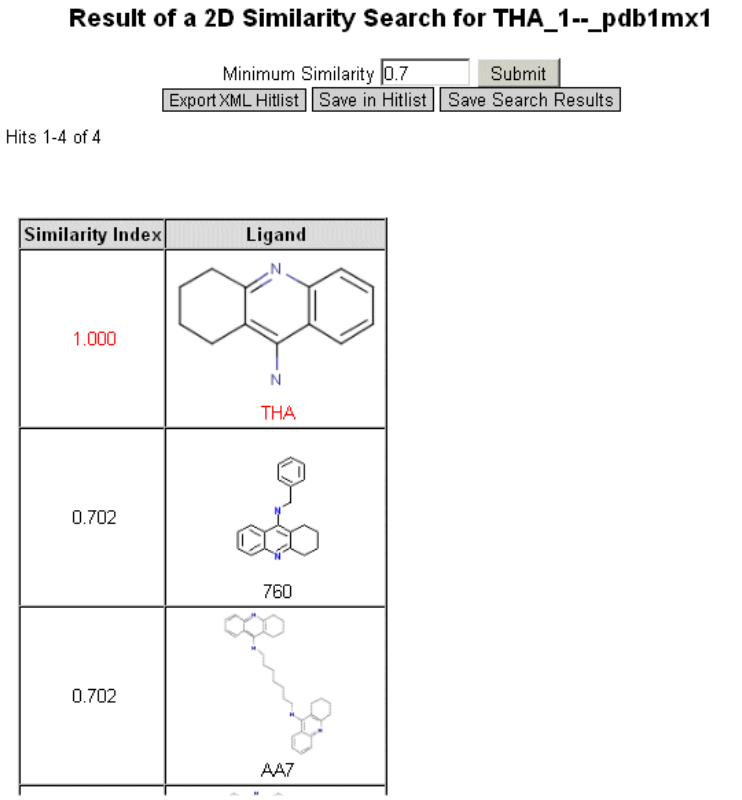
- The maximum number of hits can limited to a user-defined number. Enter the required number in the *Show maximum of [] hits* box.
- Alternatively, all hits will be returned if the *Show all hits* radio button is selected.

7 Similar Ligand Searches

- Similar ligand searches can be carried out on ligands stored in the Relibase+ database i.e. ligands from the PDB (see Searching for Similar Ligands in the PDB Section 7.1, page 78) or on ligands stored in the Cambridge Structural Database (CSD) (see Searching for Similar Ligands in the CSD Section 7.2, page 80). Note that the latter is available for CSD System subscribers only, please contact *admin@ccdc.cam.ac.uk* for further information.

7.1 Searching for Similar Ligands in the PDB

- On any ligand page, click on the **Similar Ligands** button on the menu bar above the 2D ligand diagram.
- All ligands in the Relibase+ database are compared to the reference ligand.
- The results are loaded into the browser as a list of ligands and are ranked in decreasing order of similarity to the reference ligand.
- Only the 1000 most similar results to the query ligand are returned.

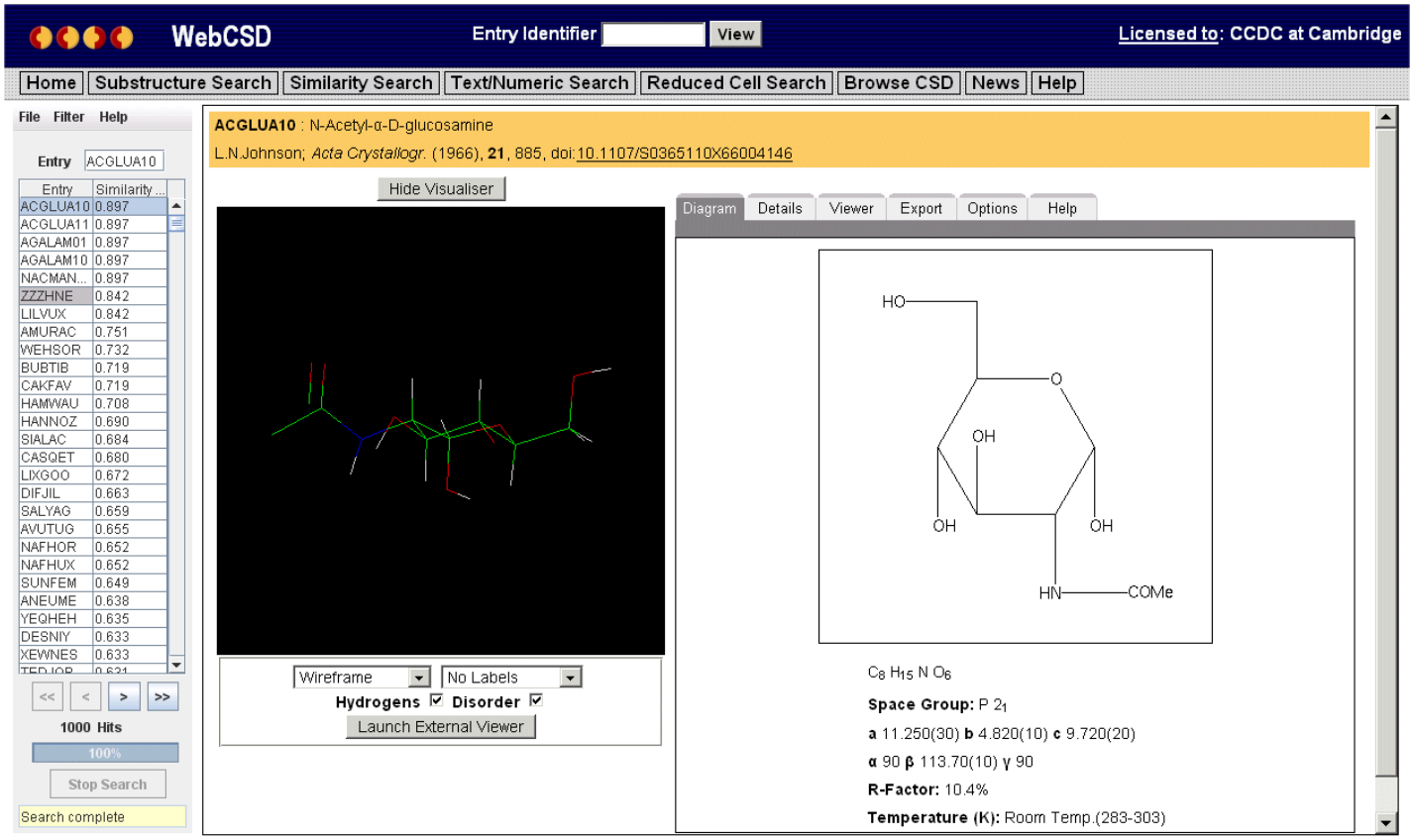


- The similarity index given in the first column is a Tanimoto coefficient. Tanimoto coefficients are calculated from a comparison of topological fingerprints of each ligand against that of the reference ligand. A fingerprint is calculated for each ligand in Relibase+ by traversing each path of up to 10 atoms within the atomic graph. At each atom in the path, a standard hashing algorithm is then used to set 2 bits in a fingerprint of 2000 bits in length. The first bit is derived from a hash code that accounts for elemental type of the current node and the and the path already traversed. The second bit is derived from a hash code that only accounts for atom types traversed along the current path. The Tanimoto coefficient is set to a default value of 0.7. During a similar ligand search, only ligands with a Tanimoto coefficient (relative to the reference ligand) above this threshold value will be displayed.
- The 2D diagrams in the second column are linked to the corresponding ligand pages.
- The search results can be filtered on the basis of the Tanimoto coefficient. Enter the required minimum similarity index (a value between 0 and 1, the default value is 0.7) into the *Minimum Similarity* window and hit the **Submit** button.
- Output options are available:
  - **Export XML Hitlist**: use this button to save the hitlist of ligand entry codes as an *XML* format file.
    Note: any saved XML hitlist can be read back into Relibase+ (see Loading XML Format Hitlists Section 6.8, page 55).

- **Save in Hitlist**: use this button to save a hitlist of ligand entry codes onto the Relibase+ server.

7.2 Searching for Similar Ligands in the CSD

- On any ligand page, click on the **Similar Ligands in CSD** button on the menu bar above the 2D ligand diagram.
- All ligands in the CSD are compared to the reference ligand.
- The results are loaded into WebCSD, the online interface to the CSD (*http://www.ccdc.cam.ac.uk/products/csd_system/webcsd/*). Only the 1000 most similar results to the query ligand are returned.



- The similarity index given in the left hand column adjacent to the 6 or 8 character CSD identifier (*refcode*) is a Tanimoto coefficient. Tanimoto coefficients are calculated from a comparison of topological fingerprints of each ligand against that of the reference ligand. A fingerprint is calculated for each ligand in the CSD by traversing each path of up to 10 atoms within the atomic graph. At each atom in the path, a standard hashing algorithm is then used to set 2 bits in a fingerprint of 2000 bits in length. The first bit is derived from a hash code that accounts for elemental type of the current node and the and the path already traversed. The second bit is derived from a hash code that only accounts for atom types traversed along the current path. The Tanimoto coefficient is set to a default value of 0.3. During a similar ligand search, only ligands

with a Tanimoto coefficient (relative to the reference ligand) above this threshold value will be displayed.

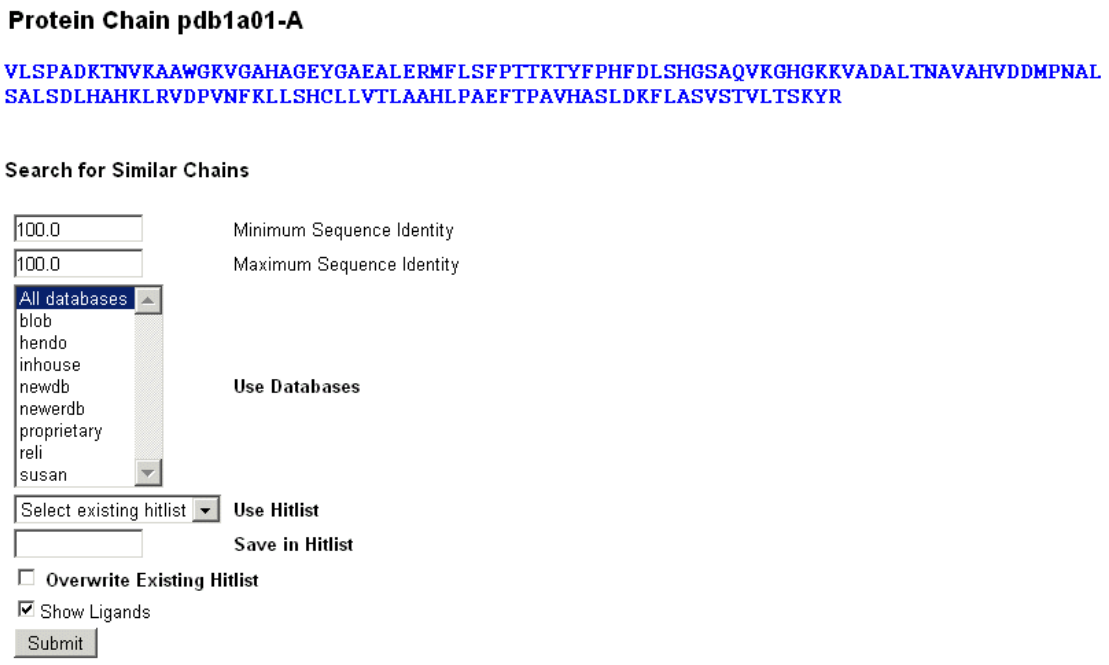
- By default, the most similar ligand (i.e. the one at the top of the hitlist) is displayed. Other ligands can be displayed by clicking on the CSD refcode e.g. **AABHTZ, AACANI10**.
- The search results can be ordered highest to lowest similarity or vice versa by clicking on the **Similarity** tab at the top of the list of similar ligands.
- A 2D diagram is provided and can be enlarged by clicking on the image.
- Further information can be accessed via the following tabs above the 2D diagram:
  - **Diagram**: the tabbed view that contains the 2D diagram and basic crystallographic information and is shown by default when the search results are loaded.
  - **Details**: this view provides more comprehensive textual information including the publication details and more comprehensive crystallographic information.
  - **Viewer**: use this tab to configure the 3D viewer size and background colour.
  - **Export**: use this tab to output the structure currently on display as either a *CIF*, an *SDFile* or a *Mol2* file.
  - **Options**: use this tab to configure the 2D diagram display options.
  - **Help**: use this tab to access help on how to use the 3D viewer.
- The 3D viewer provided is AstexViewer (see AstexViewerTM Section 5.1, page 42). Use the **Hide Visualiser** button to control whether or not the 3D view is shown. Further basic options for controlling the display are provided at the bottom of the viewer:
  - Display style: use the pulldown menu that reads *Wireframe* to pick from *Wireframe*, *Capped Sticks*, *Ball and Stick* and *Spacefill* display modes.
  - Display of labels: use the pulldown menu that reads *No labels* to show labels for *Selected* atoms, *All but C/H*, *All but C/H/N/O*, *All Metals* or *All Atoms*.
  - **Hydrogens** tickbox: use this to control whether or not H atoms are displayed (if present on the CSD structure).
  - **Disorder** tickbox: use this to control whether or not disordered atoms are displayed.
  - Use the Launch External Viewer button to view the structure in another visualiser.

8 Similar Protein Chain Searches

- On all Relibase+ entry pages, the protein chains are listed at the bottom of the protein and ligand information chart.

Protein Chains	pdb1a01-A pdb1a01-B pdb1a01-C pdb1a01-D
Nucleic Acids	No nucleic acids for this entry.
Secondary Structure	Secondary Structure Information
Waters	Water Information
Cavities	Cavity Information

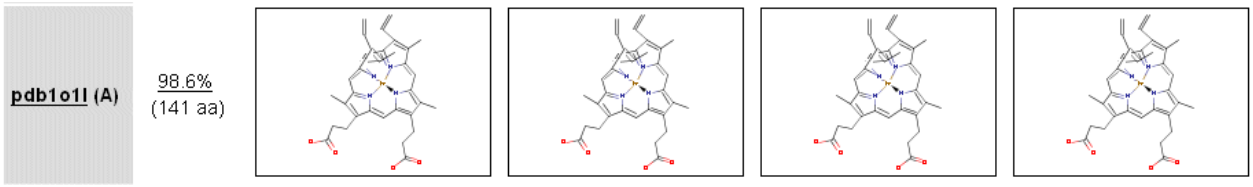
- The different sequences can be displayed by clicking on the hyperlink for the chain of interest, e.g. *pdb1a01-A* above.
- Clicking on the sequence hyperlink launches the *Protein Chain Sequence* page with a sequence display and, under this, a sequence search form.
- Forms of the type shown below are the starting point for a similar chain search, using one chain in the entry as a reference.



- *Minimum Sequence Identity* and *Maximum Sequence Identity* boxes can be used to specify the required sequence identity as a percentage with respect to the reference chain (default is 100%).
- *Use Databases* allows you to select which database or combination of databases is searched. The default setting is to search *All databases*.
- *Use Hitlist* allows you to to speed up the search by restricting it to a previously identified list

which can be selected from the pop-up menu next to *Use Hitlist*. The default is *Select existing hitlist*; until a hitlist is selected, the entire set of databases will be searched.

- *Save in Hitlist* allows you to save the results of a search in a hitlist; type the required hitlist name into the *Save in Hitlist* box before you start the search. Activate the *Overwrite Existing Hitlist* tick box if you wish to overwrite a hitlist already in existence.
  Note: it is not possible to save a hitlist of sequence similarity search results after the search has been run, however search results can be stored (see Storing Search Results Section 6.1, page 47).
- If you wish to display the ligand 2D chemical diagrams in the resulting list of chains, select the **Show Ligands** check box.
  Note: selecting this option may adversely affect the search speed.
- Hit the **Submit** button next to *Search for Similar Chains* to start the search.
- The results are displayed as a table of chains, ranked according to their sequence identity relative to the reference chain.



- The links from the % sequence identity in this table will show the alignment of two complete chains, e.g.

Alignment of Entries pdb1o1l-A and pdb1a01-A

pdb1o1l-A: 283 Amino Acids
pdb1a01-A: 141 Amino Acids
Identity: 98.6%

```
LPAEFTPAVHASLDKFLASVSTVLTSKYRGVLSPADKTNVKAAWGKVGAHAGEYGAEAWE
VLSPADKTNVKAAWGKVGAHAGEYGAEALE
RMFLSFPTTKTYFPHFDLSHGSAQVKGQGKKVADALTNAVAHVDDMPNALSALSDLHAHK
RMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHK
LRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
LRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

This alignment has been carried out using FASTA (3.5 Sept 2006) [optimized]

- Conserved residues are coloured blue, residues that are similar are coloured red, and residues that are completely different are coloured black.
- To compare the reference chain to another specific protein chain, use the Protein Chain Alignment option:

- Enter the PDB chain identifier (i.e. 1a01-B) and hit **Align**. The % sequence identity is presented for the two selected chains and the two complete chains are aligned. As before conserved residues are coloured blue, residues that are similar are coloured red, and residues that are completely different are coloured black.
  Note: the PDB code part of the chain e.g. 1a01 is ***not*** case sensitive, however the identifier (e.g. -A,) ***is*** case sensitive. If an entry has >26 chains, it is assigned an uppercase chain identifier (-A, -B etc), however if the entry has <26 chains it is assigned a lowercase chain identifier (-a, -b).

8.1 Hints for Similar Chain Searching

- Similar chain searching is recommended for retrieving a complete list of structures for a protein. For example a sequence search for thrombin will ensure that only thrombin is retrieved as a hit whereas a keyword search will retrieve many other proteins that are linked to thrombin either in structure or in biochemical function. Bear in mind that some structures of a particular protein may suffer deletions due to poor resolution of one or other loop of residues. So it may be necessary to set the lower limit of sequence identity to be less than 100%, in order to collect all structures of a particular protein.

9 Similar Binding Site Searches (and Superposition)

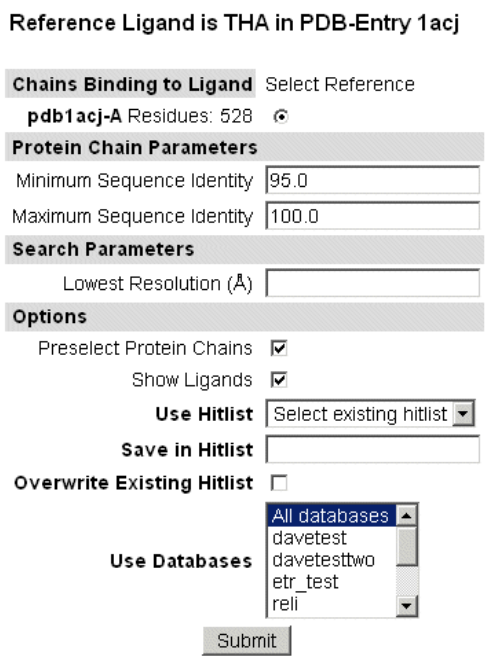
The definition of a Similar Binding Site, in the context of this section, is a binding site which has a significant degree of homology with the reference binding site. Similar binding sites in terms of surface shape and properties, but which have low homology with the reference structure, may also exist. These can be sought using the Cavbase module (see CHAPTER 4: RUNNING SIMILAR CAVITY SEARCHES, page 97). You may wish to use a similar binding site search for some of the following reasons:

- To compare the binding modes of different ligands at a particular binding site.
- To compare the binding mode of one ligand in two closely related binding sites.
- To find bioisosteric replacements.
- To analyse ligand-induced fit and protein flexibility.
- To find conserved and displaced water molecules.

9.1 Performing a Similar Binding Site Search

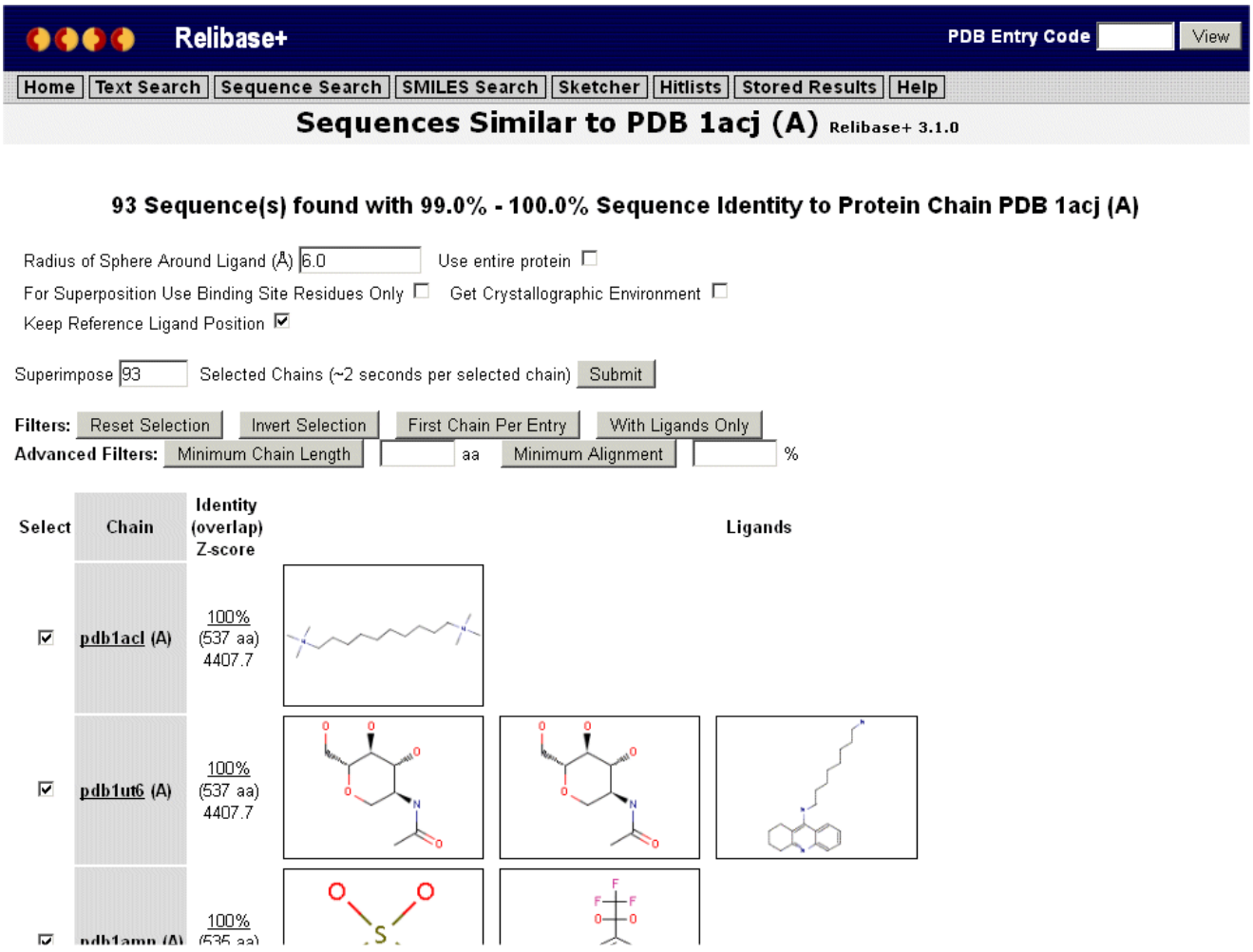
- From any ligand page, click on the **Similar Binding Sites Search** button on the menu bar above the 2D diagram of the ligand.

- A form is loaded into the browser:



- Using the radio buttons next to the chain identifiers, select the chain you wish to use as the reference chain.
- The reference chain is the chain that will be used for the sequence alignment and for the 3D superposition.
- If required, change the sequence identity limits using the *Maximum Sequence Identity* and *Minimum Sequence Identity* text boxes.
- If required, enter a resolution limit (e.g. 2.0Å) into the *Lowest Resolution* text box.
- Further options are available:
  - If you wish all chains included in the 3D superposition to be preselected in the list of results ensure that the **Preselect Protein Chains** check box is switched on (this is the default), otherwise switch off this check box and then make your selection from the list of results.
  - If you want the ligand diagrams to be displayed in the resulting table of chains, make sure the **Show Ligands** check box is selected.
  - *Use Hitlist* allows you to to restrict the search to a previously saved hitlist. Select the hitlist name from the *Use Hitlist* pulldown menu. The default is *Select existing hitlist*; until a hitlist is selected, the entire set of databases will be searched. Use the *Save in Hitlist* option to save the similar binding site search; activate the *Overwrite Existing Hitlist* tick box to overwrite an existing hitlist.
  - Similarly *Use Databases* allows you to select which database or combination of databases is searched. The default setting is to search *All databases*.

- Start the search for similar chains by clicking on the **Submit** button.
- The results are displayed as a table of chains, ranked according to their sequence identity relative to the reference chain and their Smith-Waterman score (this is a combined measure taking into account the alignment identity % and the longest sequence of matched amino acids e.g. in the screenshot below the alignment identity of the top entry is 100%, the sequence length is 537 amino acids and the Smith-Waterman score is 4407.7).
- ***Note***: the Fasta sequence search finds the 1000 best sequence matches in the sequence files, then filters the resulting chains on homology. When searching for low homologies, the required chains may not be within the 1000 best sequence matches found by Fasta.



- The links from the percentage sequence identity in this table will show the alignment of two complete chains.
- Use the check boxes in the left-hand column of the table to select or deselect the chains for 3D superposition. If you switched on the **Preselect Protein Chains** check box prior to searching, all chains will be selected automatically. Various options are available for chain selection:
  - **Reset Selection** returns you to the original chain selection, i.e that chosen for initial viewing of

the results.

- **Invert Selection** allows you to toggle back and forth between selected and deselected chains in the results list.
- **First Chain Per Entry:** only the first chain is used for superposition if the protein contains more than one chain, from your current selection.
- **With Ligands Only:** selecting this allows you to exclude entries with no ligands from your current selection.
- **Minimum Chain Length:** restricts the chain length to the value equal to or above that entered in the text box, from your current selection.
- **Minimum Alignment:** restricts the alignment that is required between the reference and superimposed chains to be equal or above a certain value.

- By default, all conserved residues in the chains are used initially for superposition, then 40% are removed for the final superposition of hits. The remaining 60% of residues are designated the 'Core' residues. If you want to use these 60% of residues for the superposition then click on the **Use Entire Protein** check box**.** The superposition algorithm uses only the alpha carbon of each residue to carry out the superposition.
- If, on the other hand, you only want to use the binding site residues from the chain, in order to carry out the chain superposition, make sure that the **For Superposition Use Binding Site Residues Only** check box is selected. Again it is the residue alpha carbons that are used in the superposition.
- If you want to look at the crystal packing for any of these similar binding sites then ensure that the **Get Crystallographic Environment** check box is selected; **Packing** buttons will then be present in the *Protein Explorer* section of the Visualiser (please refer to the Hermes documentation for further information).
- If required, adjust the radius of the sphere around the ligand by entering a new value into the *Radius of Sphere Around Ligand* text box at the top of the form (default radius is 6.0Å). If **For Superposition Use Binding Site Residues Only** is selected, this choice of radius affects the residues used for the superposition. In all cases it controls not only how much will be displayed in the 3D visualiser and but also what is used for superposition/RMS analysis (see The Analysis Table: RMS Section 9.4.2, page 91).
- Activate the *Keep Reference Ligand Position* tickbox so that the similar binding sites are superimposed on the reference ligand's original 3D coordinates (if this tick box is de-activated, the reference ligand will be moved to the origin and all binding sites will be superimposed relative to this position).
- Click on the **Submit** button to superimpose the chains and assess protein flexibility, conserved water molecules etc.

The results of a similar binding site search are presented in an analysis table. In addition all the entries in the table are displayed in AstexViewer in their superimposed states (see Appearance of

AstexViewer on a Binding Site Superposition Page Section 5.1.3, page 45). Further analysis of the superposition can be carried out using Hermes (see Analysing the results in the Hermes Visualiser Section 9.3, page 88). The entries in the analysis table will either be whole protein chains or just the binding sites and ligands depending on which option has previously been selected.

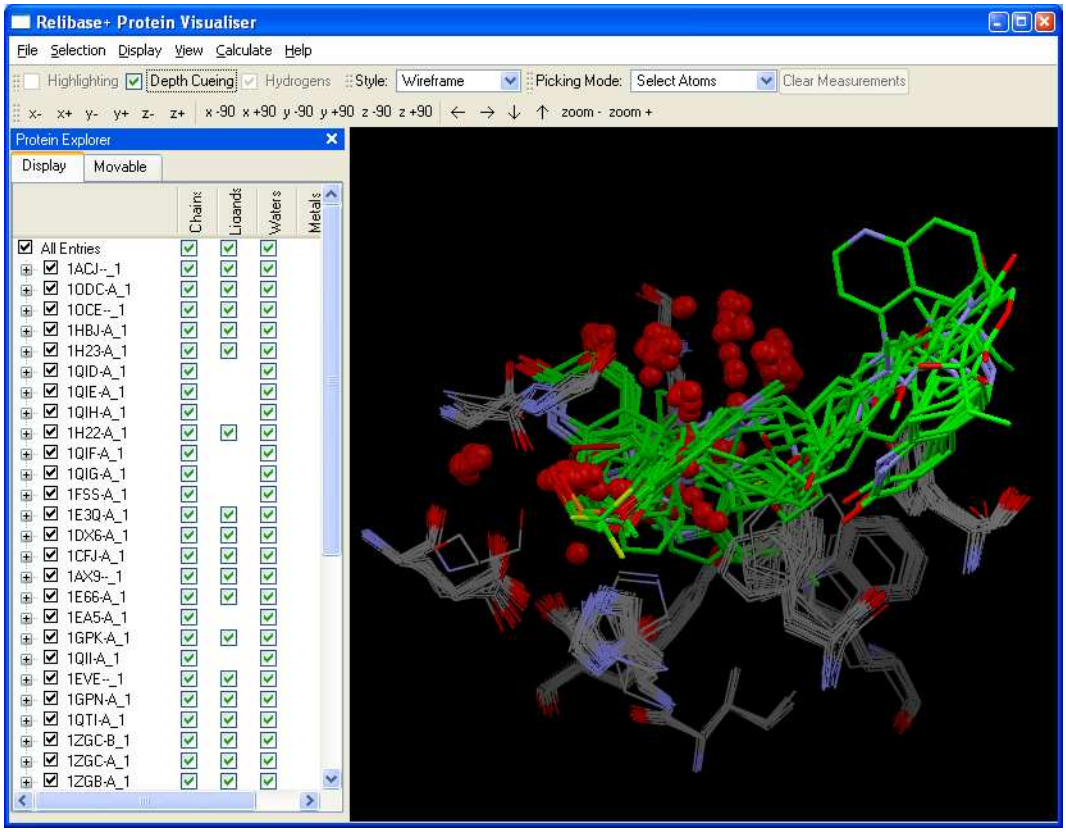
9.2 Saving Similar Binding Site Searches

Superposition searches can be saved as hitlists or the search results stored.

- To store the search results:
  - Go to the bottom of the results page and enter a name for the stored search in the *Save Superposition Results* area. Add a description of the superposition if necessary, and then click on **Save.** You will not be allowed to overwrite existing saved superposition results unless you click on the appropriate toggle box.
  - Superposition results can be retrieved via the **Stored Results** button on the menu bar at the top of the Relibase+ page. Stored superpositions may also be deleted on this page.
- To save a hitlist:
  - Go to the bottom of the results page and enter a name for the hitlist in the *Save Hitlist* area, then hit **Save**. You will not be allowed to overwrite an existing hitlist unless you click on the appropriate toggle box.

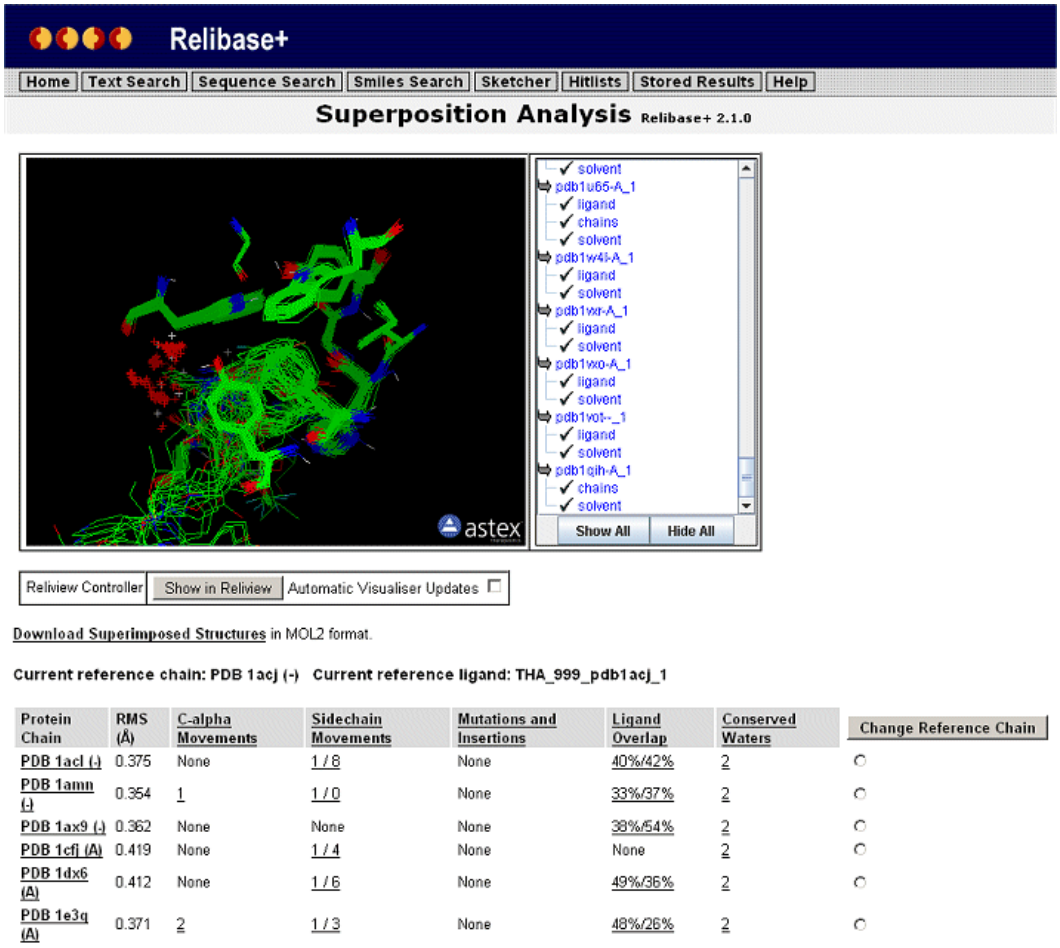
9.3 Analysing the results in the Hermes Visualiser

- For information on using *Hermes* please refer to the Hermes documentation (follow the *Hermes* link on the top right of this document).



9.4 Analysis of Superimposed Proteins/Binding Sites

A detailed analysis of the superimposed binding sites is shown in the analysis table. Information is provided on backbone and side chain movements in the protein, ligand overlap, conserved water molecules etc.:



- Search results can be viewed in AstexViewer, or if the *Automatic Visualiser Updates* tick box is activated, Hermes will automatically come up when a superposition is completed.
- It is possible to download mol2 files of all the structures in their superimposed frame of reference by hitting the **Download Superimposed structures** link. Ligands, protein chains and waters are all downloaded.
- The *current reference ligand* and *current reference chain* are displayed above the table. If you want to recalculate the superposition with a different reference chain, this can be done by clicking on the check-box next to the desired reference chain, and pressing the **Change reference chain** button. All the values in the table will be recalculated accordingly.
- There are several headers in the analysis table:
  - *Protein Chain* (see The Analysis Table: Protein Chain Section 9.4.1, page 91)
  - *RMS* (see The Analysis Table: RMS Section 9.4.2, page 91)
  - *C-Alpha Movements* (see The Analysis Table: C-Alpha Movements Section 9.4.3, page 91)
  - *Sidechain Movements* (see The Analysis Table: Sidechain Movements Section 9.4.4, page 92)
  - *Mutations and Insertions* (see The Analysis Table: Mutations and Insertions Section 9.4.5, page 93)

- *Ligand Overlap* (see The Analysis Table: Ligand Overlap Section 9.4.6, page 94)
- *Conserved Waters* (see The Analysis Table: Conserved Waters Section 9.4.7, page 95)
- *Clashes with Proteins* (see The Analysis Table: Clashes with Proteins Section 9.4.8, page 95)

9.4.1 The Analysis Table: Protein Chain

- The first column presents the superimposing protein chain. Each entry in this column links to a PDB entry page.

9.4.2 The Analysis Table: RMS

- The second column gives the RMS figure for the entire chain-on-chain superposition, *RMS(overall)*. The RMS is calculated from the alpha carbons of each correctly aligned residue. This is the RMS value that is given by default.
- Two additional RMS values can also be calculated by checking the *RMS(core)* and *RMS(binding site)* check boxes at the base of the table; and then clicking on the **Recalculate Table** button, again at the base of the table.
  - The *RMS(core)* is the RMS calculated using only the alpha carbons of the *Core* residues. The core residues are those residues selected for final superposition of hits after the initial superposition of all conserved residues is performed (see Performing a Similar Binding Site Search Section 9.1, page 84).
  - The *RMS(binding site)* is the RMS calculated using the alpha carbons that make up the binding site defined using the option *Radius of Sphere Around Ligand (Å)* in the similar binding sites superposition setup page (default value is 6Å) (see Performing a Similar Binding Site Search Section 9.1, page 84).

9.4.3 The Analysis Table: C-Alpha Movements

- The third column gives information on significant C-alpha movements (if any) with respect to the reference chain. The default threshold for what constitutes a significant movement is 0.5Å. This threshold can be changed by altering the relevant figure in the *C-alpha Movements box* in the *Protein flexibility* area at the base of the table and then clicking **Recalculate Table**. The column can also be hid from view by clicking off the appropriate check box in the *Protein flexibility* area, prior to recalculation.
- Each numeric entry in the *C-Alpha Movements* column links to an expanded list of residues involved in movement, and the distance of movement in each case.

Protein chain: PDB 1e66 (A)

Movement of C-alpha atoms :

Residue	Distance / Å
SER 4	1.688
GLU 5	0.668
VAL 22	0.658
LEU 23	0.883
SER 24	0.831
SER 25	0.618
VAL 71	0.531

- The header of the third column links to a summary table of C-alpha movements for all chains in the Analysis Table. Movements of greater than 1.0Å are highlighted in red.

C-alpha Movements

Reference chain: PDB 1acj (-)
Reference ligand: THA_999_pdb1acj_1

Note: values above a threshold of 1Å are highlighted in red.

Protein Chain	SER4	GLU5	VAL22	LEU23	SER24	SER25	VAL71	PRO76	GLY77	LEU146	SER147	GLY160	GLU261	ILE347	HIS362	ALA363
PDB 1e66 (A)	1.688	0.668	0.658	0.883	0.831	0.618	0.531	0.656	0.533	0.654	0.555	0.574	0.521	0.517	0.624	0.667
PDB 1ea5 (A)	1.498	0.505	0.505	0.721	0.779	0.569	0.654			0.578	0.518		0.522		0.781	0.734
PDB 1gpk (A)	1.380		0.680	0.801	0.924	0.678	0.507		0.587	0.578	0.606				0.775	0.755
PDB 1gqr (A)	1.315			0.851		0.581	0.740			0.523	0.556	0.620		0.620	1.401	0.648
PDB 1h22 (A)	1.920	0.553	0.590	0.708	1.091	0.584		0.768	0.682	0.548	0.531		0.511		0.883	0.883
PDB 1h23 (A)	1.513	0.521	0.501	0.700	1.132	0.502		0.708	0.519	0.585	0.547	0.537	0.500		0.774	0.916
PDB 1odc (A)	1.815	0.868	0.606	0.719	1.243	0.532		0.725	0.809	0.541	0.554	0.618	0.572	0.504	0.966	0.987
PDB 1qid (A)	1.482	0.617	0.570	0.771	0.773	0.587	0.668			0.581	0.555	0.531		0.513	0.726	0.656

9.4.4 The Analysis Table: Sidechain Movements

- The fourth column gives information on the number of significant sidechain movements (if any) with respect to the reference chain (first figure). The movement is measured between the centroids calculated for all the heavy atoms within the sidechain, for both reference and superimposed chains. Also given are the number of sidechain torsion angles that differ significantly from those in the reference chain (second figure). The default threshold for what constitutes a significant atom movement is 1.0Å.The default threshold for what constitutes a torsion change, is 10 degrees. These thresholds can be changed by altering the relevant figures in the *Sidechain movements* and *Torsion angle changes box* in the *Protein flexibility* area at the base of the table, and then clicking **Recalculate Table**. The column can also be hid from view by clicking off the appropriate check box in the *Protein flexibility* area, prior to recalculation.
- Each numeric entry in the *Sidechain Movements* column links to an expanded list of residues involved in movement, and the distance of sidechain centre movement in each case. Below this list are tabulated details of the significantly different torsions that have been identified.

.

Protein chain: PDB 1gqr (A)

Movement of side chain centers :

Residue	Distance / Å
PHE 330	2.340
HIS 440	1.148

Changes in torsion angles :

Residue	Torsion Atoms	Reference Torsion / °	New Torsion / °	Difference / °
TRP 84	CA-CB-CG-CD1	98.178	112.135	13.957
PHE 330	C-CA-CB-CG	31.387	107.844	76.457
PHE 330	CA-CB-CG-CD1	61.773	117.354	55.581
HIS 440	C-CA-CB-CG	61.866	0.285	61.580
TYR 442	C-CA-CB-CG	138.699	149.446	10.747

- The header of the fourth column links to a summary table of sidechain movements for all chains in the Analysis Table. Movements of greater than 1.5Å are highlighted in red.

Sidechain Movements

Reference chain: PDB 1acj (-)
Reference ligand: THA_999_pdb1acj_1

Note: values above a threshold of 1.5Å are highlighted in red.

Protein Chain	PHE330	HIS440
PDB 1e66 (A)		
PDB 1ea5 (A)	2.476	
PDB 1gpk (A)	1.183	
PDB 1gqr (A)	2.340	1.148
PDB 1h22 (A)	2.249	
PDB 1h23 (A)	2.299	

9.4.5 The Analysis Table: Mutations and Insertions

- The fifth column tabulates the total number of mutations/insertions that occur between the reference chain and the superimposed chain. As before, information is only tabulated for those regions defined by the user i.e. either the whole protein, or the binding site as defined by radius from the ligand. The column can be hidden from view in subsequent recalculations, by clicking off the relevant checkbox in the *Protein flexibility* area at the base of the table.
- Each numeric entry in the *Mutations and Insertions* column links to an expanded list that provides further information.

Mutations and Insertions

Reference chain: PDB 1fax (A)
Reference ligand: DX9_1_pdb1fax_1

Protein chain: PDB 1bbr (E)

Mutations and Insertions :

Reference Residue	Mutated to Superimposed Residue
PHE 174	ILE 174
GLN 192	GLU 192
ALA 221	ASP 221
ILE 227	PHE 227

Insertion in reference GLU 97
Insertion in reference THR 98

- Clicking on the header to column five links to a concatenation of the mutation/insertion data for each relevant chain. All chains are represented.

9.4.6 The Analysis Table: Ligand Overlap

- The sixth column in the table gives the percent of ligand overlap between the ligands in the reference and superimposed chains. The first figure is the percent overlap in terms of reference ligand volume, the second figure is the percent overlap in terms of superimposed ligand volume. The reference ligand is always the ligand that was originally used to set up the superposition analysis.To change the reference ligand it is necessary to start from a different ligand page.The column can be hidden from view in subsequent recalculations, by clicking off the relevant checkbox in the *Ligand/binding site* area at the base of the table.
- Each entry in the sixth column links to an expansion of the information available in the table. The ligand pages of both reference and superimposed ligands can be accessed from here.
- Clicking on the header to column six links to a concatenation of the ligand overlap data for each chain. All relevant chains are represented.

Protein chain: PDB 1bbr (E)

Overlap of ligands :

Superimposed Ligand	Volume of Reference / %	Reference Ligand	Volume of Superimposed / %
ACE-DFLAEGGGV_308-F_pdb1bbr_1	60.618	DX9_1_pdb1fax_1	33.17

Protein chain: PDB 1bhx (F)

Overlap of ligands :

Superimposed Ligand	Volume of Reference / %	Reference Ligand	Volume of Superimposed / %
R56_1-L_pdb1bhx_1	59.113	DX9_1_pdb1fax_1	73.424

9.4.7 The Analysis Table: Conserved Waters

- The seventh column of the table gives the number of conserved waters that have been identified within the region of the protein under consideration. A conserved water is defined as one that is within 1.2Å of a water in the reference binding site, after superposition. The column can be hidden from view in subsequent recalculations, by clicking off the relevant checkbox in the *Ligand/binding site* area at the base of the table.
- Each numeric entry in the seventh column links to a table that gives the residue numbers of the waters that are considered conserved in the superimposed structures. The corresponding waters in the reference structure are also identified.
- The header of column seven links to a table which tabulates, under each water in the reference structure that is relevant, the details of the corresponding conserved waters in all the superimposed chains. If a water is displaced by a ligand, in one or other of the superimposed chains, then this information is also tabulated. A link is available for the appropriate ligand page.

Conserved/Displaced Waters

Reference chain: PDB 1acj (-)
Reference ligand: THA_999_pdb1acj_1

Key: Grey = conserved; Blue = displaced

Reference Water 634 (Temperature Factor 24.060Å^2); Conserved: 10, Displaced: 1				
PDB 1odc (A) - Water 78	**Distance:**	1.103 Å	**Temperature Factor:**	35.650 Å^2
PDB 1e66 (A) - Water 157	**Distance:**	0.843 Å	**Temperature Factor:**	24.790 Å^2
PDB 1som (A) - Water 688	**Distance:**	0.903 Å	**Temperature Factor:**	34.040 Å^2
PDB 1zgc (A) - Water 60	**Distance:**	1.150 Å	**Temperature Factor:**	23.210 Å^2
PDB 1zgc (B) - Water 63	**Distance:**	1.107 Å	**Temperature Factor:**	19.970 Å^2
PDB 1w6r (A) - Water 65	**Distance:**	0.937 Å	**Temperature Factor:**	30.630 Å^2
PDB 1vxr (A) - Water 1022	**Distance:**	0.824 Å	**Temperature Factor:**	29.540 Å^2
PDB 1qie (A) - Water 1130	**Distance:**	0.865 Å	**Temperature Factor:**	38.020 Å^2
PDB 1qif (A) - Water 1164	**Distance:**	0.925 Å	**Temperature Factor:**	45.910 Å^2
PDB 1qid (A) - Water 1127	**Distance:**	0.943 Å	**Temperature Factor:**	34.040 Å^2
HUP_1540-A_pdb1qpk_1	**Atom:**	C3	**Sybyl Type:**	C.2
Reference Water 643 (Temperature Factor 20.340Å^2); Conserved: 14, Displaced: 0				
PDB 1h22 (A) - Water 36	**Distance:**	0.484 Å	**Temperature Factor:**	29.840 Å^2
PDB 1h23 (A) - Water 44	**Distance:**	0.416 Å	**Temperature Factor:**	31.440 Å^2

9.4.8 The Analysis Table: Clashes with Proteins

- This data is not presented in the analysis table as it is first calculated. Click on the relevant checkbox in the *Ligand/binding site* area at the base of the table and then click on the **Recalculate Table** button. Each number in the column represents the number of atoms in the reference ligand which clash with atoms in the relevant superimposed chain. A clash is defined as being an atom-atom distance of less than the sum of the Van der Waal's radii by 0.1Å or more.
- Each numeric entry in the column links to a table giving further information about the clashes found for that chain.

Conserved/Displaced Waters

Reference chain: PDB 1acj (-)
Reference ligand: THA_999_pdb1acj_1

Key: Grey = conserved; Blue = displaced

Reference Water 634 (Temperature Factor 24.060Å^2); Conserved: 10, Displaced: 1				
PDB 1odc (A) - Water 78	**Distance:**	1.103 Å	**Temperature Factor:**	35.650 Å^2
PDB 1e66 (A) - Water 157	**Distance:**	0.843 Å	**Temperature Factor:**	24.790 Å^2
PDB 1som (A) - Water 688	**Distance:**	0.903 Å	**Temperature Factor:**	34.040 Å^2
PDB 1zgc (A) - Water 60	**Distance:**	1.150 Å	**Temperature Factor:**	23.210 Å^2
PDB 1zgc (B) - Water 63	**Distance:**	1.107 Å	**Temperature Factor:**	19.970 Å^2
PDB 1w6r (A) - Water 65	**Distance:**	0.937 Å	**Temperature Factor:**	30.630 Å^2
PDB 1vxr (A) - Water 1022	**Distance:**	0.824 Å	**Temperature Factor:**	29.540 Å^2
PDB 1qie (A) - Water 1130	**Distance:**	0.865 Å	**Temperature Factor:**	38.020 Å^2
PDB 1qif (A) - Water 1164	**Distance:**	0.925 Å	**Temperature Factor:**	45.910 Å^2
PDB 1qid (A) - Water 1127	**Distance:**	0.943 Å	**Temperature Factor:**	34.040 Å^2
HUP_1540-A_pdb1qpk_1	**Atom:**	C3	**Sybyl Type:**	C.2
Reference Water 643 (Temperature Factor 20.340Å^2); Conserved: 14, Displaced: 0				
PDB 1h22 (A) - Water 36	**Distance:**	0.484 Å	**Temperature Factor:**	29.840 Å^2
PDB 1h23 (A) - Water 44	**Distance:**	0.416 Å	**Temperature Factor:**	31.440 Å^2

- Clicking on the header to the column links to a concatenation of the individual clash data for each relevant chain. All chains are represented.

CHAPTER 4: RUNNING SIMILAR CAVITY SEARCHES

1 Introduction and Background Theory (see page 97)
2 Accessing Cavity Information for Relibase+ Database Entries (see page 99)
3 Displaying and Comparing Cavities (see page 101)
4 Cavity Similarity Searching (see page 107)
5 Saving Cavities to File (see page 118)
6 Building In-House Cavity Databases (see page 118)

1 Introduction and Background Theory

1.1 Introduction to CavBase

- CavBase is a program that can detect unexpected similarities amongst protein cavities (e.g. active sites) that share little or no sequence homology.
- The program is supplied with a database of cavities from PDB protein structures, including, but not confined to, known small-molecule binding sites. Any cavity (or part of a cavity) from this database can be used as query in a similarity search which will find other, similar cavities or sub-cavities in the remainder of the database.
- Similarity is judged by matching 3D property descriptors (pseudocentres) that encode the shape and chemical characteristics of each cavity (see Pseudocentres Section 1.4, page 98). No sequence information is used, which is why the program can detect similar cavities even if they have no obvious secondary-structure relationship.
- Visualisation software is provided for displaying the results of cavity similarity searches, for comparing query and hit cavities, etc. (see Displaying and Comparing Cavities Section 3, page 101).
- The CavBase cavity database is closely linked to, and may be used alongside, Relibase+.
- CavBase enables cavity databases to be created from in-house protein structures (see Building In-House Cavity Databases Section 6, page 118) and searched alongside the PDB-derived database.

1.2 Uses of CavBase

The main uses of CavBase are:

- Inference of function/mechanism of active sites (by comparing the query cavity with similar cavities of known function/mechanism).
- Generation of ideas for novel ligands (by observing what is bound to other, similar sites).
- Investigation of ligand selectivity and cross-reactivity (since a ligand known to bind to the query cavity might bind to other, similar cavities).
- Identification of novel target sites (since the database contains all cavities found by the cavity-detection algorithm, not just those of known significance).

1.3 Cavity Detection

- Cavities on the protein surfaces are detected using a modified version of LIGSITE (see References, page 172), which identifies surface depressions based on a grid-based geometrical algorithm.
- Very large cavities (> 3000Å^3) are omitted from the database as they are usually of little interest (they are often ill-defined gaps between large protein domains).
- Some shallow or completely enclosed cavities are missed by the detection algorithm.

1.4 Pseudocentres

- A simple scheme is used for deriving 3D descriptors that encode the surface properties of all the cavities (see References, page 172). All amino acid residues lining a cavity are analysed to ascertain whether they determine the chemical property of the nearby surface. This is based on geometric considerations, e.g. a C=O acceptor group pointing at the surface will be assumed to confer the property *acceptor* to the patch of the surface where the oxygen atom is exposed. In contrast, chemical groups pointing away from the surface will be neglected. A dummy atom, or *pseudocentre*, is placed on the surface to represent the chemical property that is expressed by the atom(s) exposed in that area, e.g. an acceptor patch of the surface would have an acceptor pseudocentre placed upon it.
- A cavity is thus described by a set of pseudocentres, each of which is characterised by a property (currently: donor, acceptor, donor/acceptor, aromatic, aliphatic, pi and metal) and 3D-coordinates.
- The rules for encoding pseudocentre assignment are given in the paper by Schmitt *et al* (see References, page 172). Additional pseudocentres have been assigned for the Relibase+ implementation of CavBase:
  - Pi-type pseudocentres are used to represent the hydrophobic character above/below amide, guanidinium and carboxylate planes (backbone, Asn, Asp, Arg, Gln, Glu).
  - Trp indole rings are represented by two aromatic-type pseudocentres, rather than one, as given in the paper.
  - Aliphatic pseudocentres can be assigned to the side chains of all amino acids. The atoms taken into account for placing these pseudocentres do not include atoms of functional groups (such as carboxylate). For example, for Glu, the CB and CG carbon atoms are considered, but not the carboxylate CD atom.
  - Discrete metal atoms are treated as pseudocentres.
    Note that metals of cofactors and metal-containing ligands are not treated as part of the binding site and are therefore ignored by the pseudocentre-generation algorithm. Also note that there will be no *Metal* pseudocentre shown in the *Visualisation Controls* part of Hermes interface if there is no metal in the structure.

- All the information described above is precalculated and stored in a Relibase+ type database.

1.5 Similarity Searching and Scoring

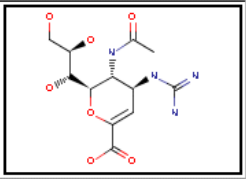
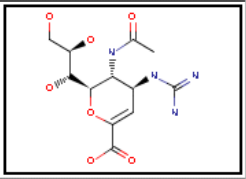
- Similarity searching in CavBase (see Cavity Similarity Searching Section 4, page 107) aims to find cavities or sub-cavities in the database (or database*s* if you have an in-house database) that match well with a query. The query must itself be a cavity or sub-cavity from the database. ***Note***: only proteins that have cavities will be searched, not the entire Relibase+ database.
- The similarity-search program employs a clique-detection algorithm, described by Schmitt et al. (see References, page 172), which treats the pseudocentres of the two cavities being compared as nodes of a graph. The algorithm effectively matches some or all pseudocentres from the query onto similar pseudocentres in the hit. The cavities are then superimposed by least-squares fitting of the pseudocentres associated with the clique-solution. Pseudocentres with short pairwise distances and matching chemical properties are extracted, and a simple function that estimates the overlap of the respective surface patches is used for calculating a similarity-score value.
- There is a choice of scoring functions. Scoring function 1 is the original scoring function, as described by Schmitt et al. (see References, page 172). Scoring functions 2 and 3 are modifications of the original function developed at CCDC, specifically 2 and 3 differ in the way that the overlap in surface points are calculated. In the original function the degree of mutual overlap is expressed by the number of surface points (from pseudocenters) within a distance threshold of 1.0Å (the points are simply counted). In 2 and 3 this simple distance function (on/off) is replaced by a more complicated more accurate block function in order to better evaluate the degree of overlap between two surface-patches. The surface points on each patch created by the pseudocenters are not only counted but also weighed according to the closest distance between two surface points - two surface points that are close give a higher contribution to the score. The difference in 2 and 3 is in the distance that is used in the block function, in scoring function 2 the distance is squared leading to a different distance dependence of the score. Scoring function 3 also has a better pair selection of pseudocenters, improving the calculation of overlap. At present, there is insufficient information regarding the relative merits of these functions; all appear to work reasonably well. The functions are on different scales, so comparisons between the values from different functions are invalid. However, for any given function, a higher number always suggests closer similarity.

2 Accessing Cavity Information for Relibase+ Database Entries

- Cavity information for any database entry can be accessed by clicking on the **Cavity Information** button at the bottom of either Protein or Ligand Information pages:

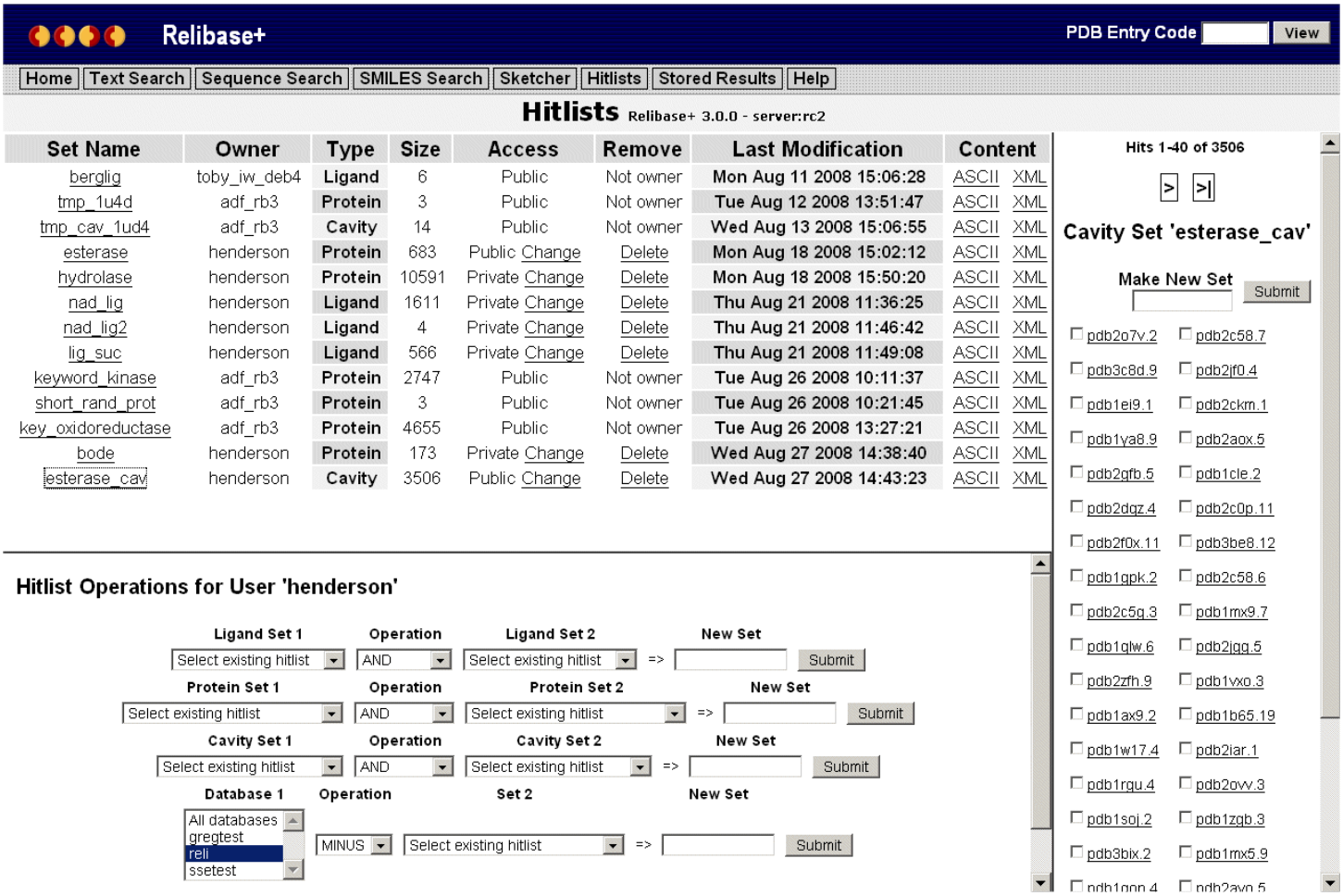
- If you are on a Ligand Information page, clicking on this button takes you to a page displaying the volume of the cavity containing the ligand, header information, and the ligand chemical diagram (if available). Also, Hermes will open with the selected cavity loaded (see Displaying and Comparing Cavities Section 3, page 101).
- If you are on a Protein Information page, clicking on the **Cavity Information** button takes you to a page listing all the cavities in the protein structure, with their volumes and any ligands they contain (clicking on a ligand diagram links you to the Ligand Information page). Selecting one of these cavities will then give you fuller details of that cavity and load it into Hermes.

Cavities in PDB Entry 1a4g

Cavity	Volume ($Å^3$)	Ligand
pdb1a4g.1	605.2	
pdb1a4g.2	689.0	
pdb1a4g.3	878.4	
pdb1a4g.4	1055.9	
pdb1a4g.5	1501.4	

Similarity Search Results

- If you have created a cavity hitlist (see Searching a Subset of the Database Section 4.4.1, page 112) and want to view one of the cavities in it, display the hitlist-manager page by clicking on the top-level **Hitlist** button, and then click on the name of relevant cavity hitlist (e.g. **esterase_cav** below). This will display all the entries in the hitlist:

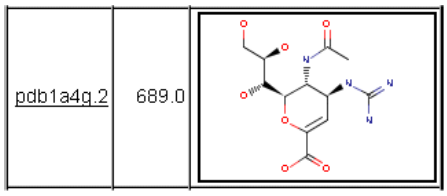

Relibase+
PDB Entry Code
View
Home
Text Search
Sequence Search
SMILES Search
Sketcher
Hitlists
Stored Results
Help
Hitlists Relibase+ 3.0.0 - server:rc2
Set Name | Owner | Type | Size | Access | Remove | Last Modification | Content
berglig | toby_iw_deb4 | Ligand | 6 | Public | Not owner | Mon Aug 11 2008 15:06:28 | ASCII XML
tmp_1u4d | adf_rb3 | Protein | 3 | Public | Not owner | Tue Aug 12 2008 13:51:47 | ASCII XML
tmp_cav_1ud4 | adf_rb3 | Cavity | 14 | Public | Not owner | Wed Aug 13 2008 15:06:55 | ASCII XML
esterase | henderson | Protein | 683 | Public Change | Delete | Mon Aug 18 2008 15:02:12 | ASCII XML
hydrolase | henderson | Protein | 10591 | Private Change | Delete | Mon Aug 18 2008 15:50:20 | ASCII XML
nad_lig | henderson | Ligand | 1611 | Private Change | Delete | Thu Aug 21 2008 11:36:25 | ASCII XML
nad_lig2 | henderson | Ligand | 4 | Private Change | Delete | Thu Aug 21 2008 11:46:42 | ASCII XML
lig_suc | henderson | Ligand | 566 | Private Change | Delete | Thu Aug 21 2008 11:49:08 | ASCII XML
keyword_kinase | adf_rb3 | Protein | 2747 | Public | Not owner | Tue Aug 26 2008 10:11:37 | ASCII XML
short_rand_prot | adf_rb3 | Protein | 3 | Public | Not owner | Tue Aug 26 2008 10:21:45 | ASCII XML
key_oxidoreductase | adf_rb3 | Protein | 4655 | Public | Not owner | Tue Aug 26 2008 13:27:21 | ASCII XML
bode | henderson | Protein | 173 | Private Change | Delete | Wed Aug 27 2008 14:38:40 | ASCII XML
esterase_cav | henderson | Cavity | 3506 | Public Change | Delete | Wed Aug 27 2008 14:43:23 | ASCII XML
Hits 1-40 of 3506
Cavity Set 'esterase_cav'
Make New Set
Submit
pdb2o7v.2
pdb2c58.7
pdb3c8d.9
pdb2jf0.4
pdb1ei9.1
pdb2ckm.1
pdb1ya8.9
pdb2aox.5
pdb2qfb.5
pdb1cle.2
pdb2dqz.4
pdb2c0p.11
pdb2f0x.11
pdb3be8.12
pdb1gpk.2
pdb2c58.6
pdb2c5g.3
pdb1mx9.7
pdb1qlw.6
pdb2jgg.5
pdb2zfh.9
pdb1vxo.3
pdb1ax9.2
pdb1b65.19
pdb1w17.4
pdb2iar.1
pdb1rgu.4
pdb2ovv.3
pdb1soj.2
pdb1zgb.3
pdb3bix.2
pdb1mx5.9
Hitlist Operations for User 'henderson'
Ligand Set 1
Operation
Ligand Set 2
New Set
Select existing hitlist
AND
Select existing hitlist
=>
Submit
Protein Set 1
Operation
Protein Set 2
New Set
Cavity Set 1
Operation
Cavity Set 2
New Set
Database 1
Operation
Set 2
New Set
All databases
gregtest
reli
ssetest
MINUS

- From the resulting list of cavities, select the hyperlink next to the check box for the cavity you wish to view.

3 Displaying and Comparing Cavities

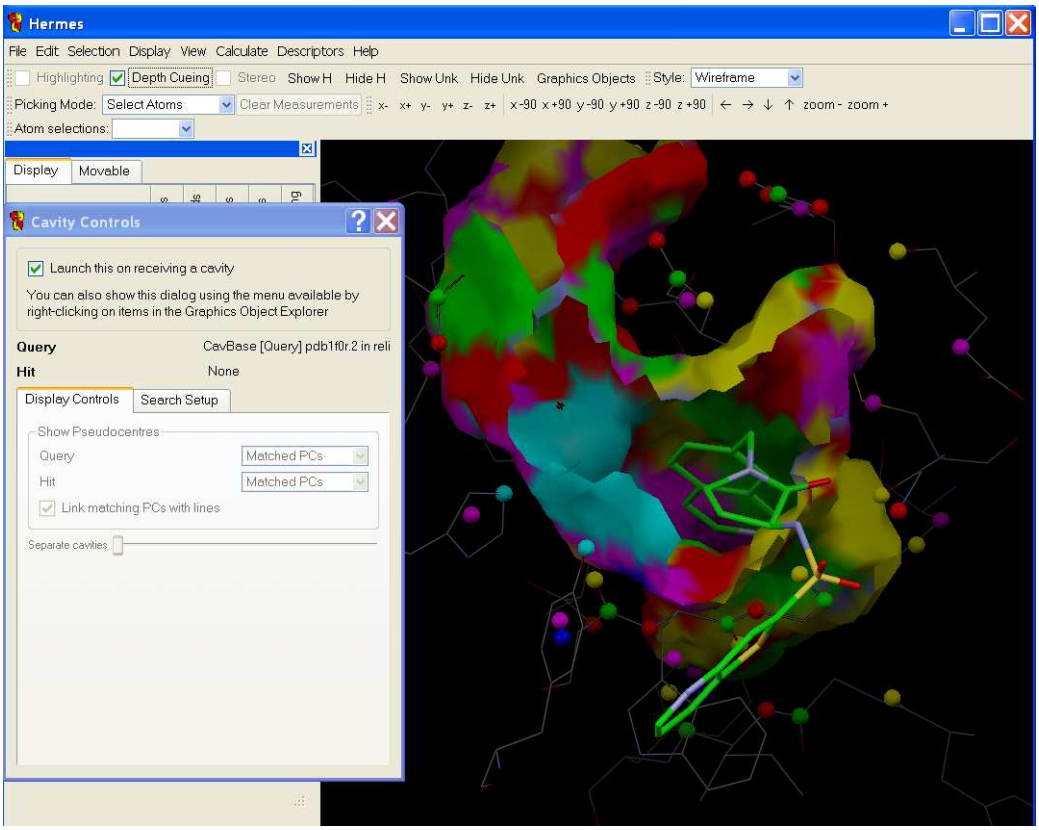
3.1 Displaying a Cavity

- A view of the cavity opens automatically in Hermes whenever you click on a cavity hyperlink, e.g. the text *pdb1a4g.2* in the display below:



- Further details of how to access cavity information are given elsewhere (see Accessing Cavity Information for Relibase+ Database Entries Section 2, page 99).

- The protein around the cavity and any ligands bound will be displayed in the visualiser. The cavity itself will appear as a solid surface which is coloured according to the aadjacent pseudocentre types. The correspoding pseudocentres are also displayed. In addition a *Cavity Controls* window will appear.


Hermes
File Edit Selection Display View Calculate Descriptors Help
Highlighting Depth Cueing Stereo Show H Hide H Show Unk Hide Unk Graphics Objects Style: Wireframe
Picking Mode: Select Atoms Clear Measurements
Atom selections:
Display Movable
Cavity Controls
Launch this on receiving a cavity
You can also show this dialog using the menu available by right-clicking on items in the Graphics Object Explorer
Query CavBase [Query] pdb1f0r.2 in reli
Hit None
Display Controls Search Setup
Show Pseudocentres
Query Matched PCs
Hit Matched PCs
Link matching PCs with lines
Separate cavities

- Some instructions on how to manipulate cavities and initiate cavity searches are given in the following sections. For more information on viewing, modifying and manipulating structures and other items in the Hermes visualiser please refer to the relevant section of the manual (follow the *Hermes* link on the top right of this document).

3.2 Moving Objects in the 3D Display

3.2.1 Rotating, Translating and Scaling

- The 3D display can be rotated by moving the cursor in the display area while keeping the left-hand mouse button pressed down (*x* and *y* rotation), or while keeping both the left-hand mouse button and the *Shift* key pressed down (*z* rotation).
- If you have a mouse with three buttons (or two buttons and a scroll wheel), the contents of the display area can be translated by moving the cursor while holding the middle mouse button (or scroll wheel) down. Alternatively, use the left-hand mouse button with the keyboard *Ctrl* key pressed down.

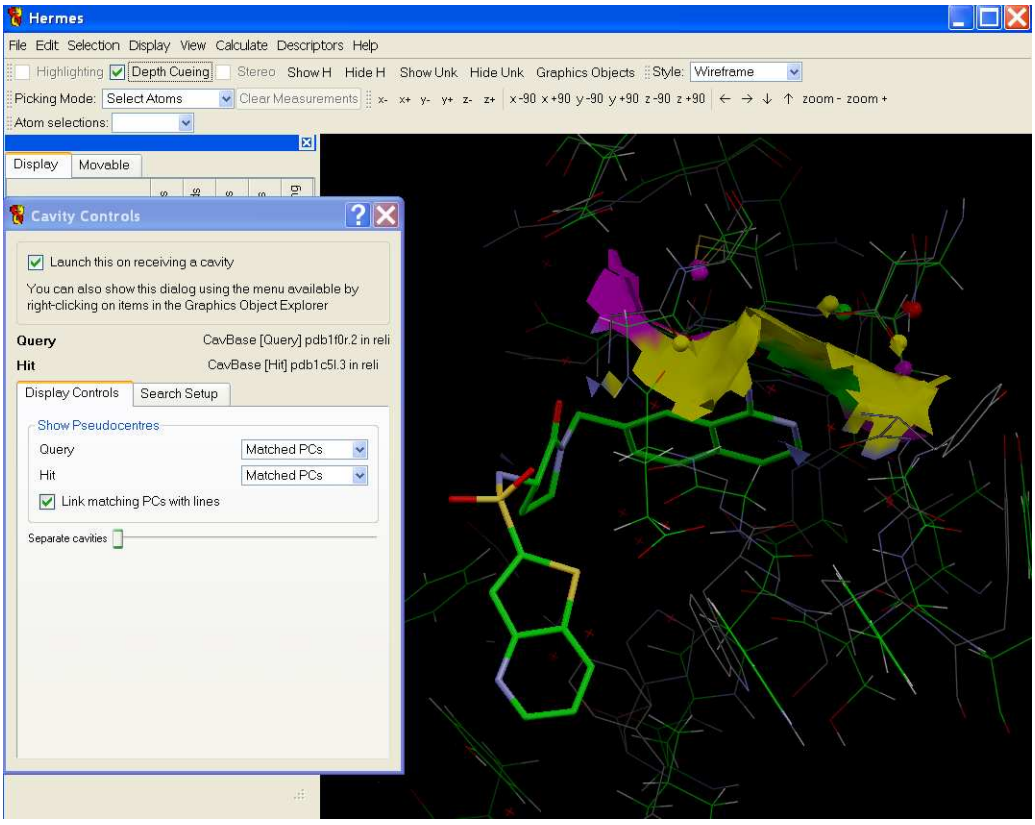
- The contents of the display area can be scaled (i.e. zoomed in or out) by moving the cursor up and down in the display area while keeping the right-hand mouse button pressed down.

3.2.2 Controlling Cavity Separation

- If you load a hit cavity from a similarity search, it will, by default, be superimposed on the query cavity. The default behaviour is for the surface areas of the query and hit cavities to only be displayed where they match. The corresponding pseudocentres will be also be displayed and superimposed. The spherical pseudocentres arise from the query whereas the tetrahedral pseudocentres arise from the hit cavity.

3.2.3 Displaying Matching and Non-matching Pseudocentres

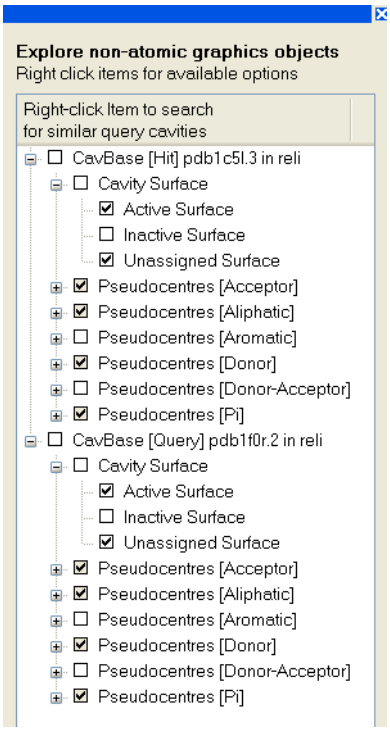
- It is possible to change which pseudocentres are displayed for hit and query cavities via the **Query** and **Hit** drop-down menus in the *Cavity Controls window.* Options in the upper menu are to display **All PC's in cavity**, **PC's searched for**, **Matched PC's** (default) and **Unmatched PC's**. Only the last two options are avaliable in the lower menu.
- The two cavities can be moved apart by using the *Separate cavities* slider bar at the bottom of the *Cavity Controls* pane. If you separate the cavities, you will probably need to zoom out as well (see Rotating, Translating and Scaling Section 3.2.1, page 102), to keep everything in view.



3.3 Customising the Cavity Display

3.3.1 Using the Graphics Object Explorer to Control the Display of Cavity Objects

- When a cavity is opened two dockable windows also become visible to the left of the Hermes screen. The upper is the *Protein Explorer Window* which can be used to control many aspects of structure display (please refer to the Hermes documentation for further information). The lower is the *Graphics Object Explorer* which is used to control the display of graphical objects which are not chemical structures. The display of cavity surfaces and pseudocentres is controlled by this window.
- Hierarchical "trees" are used to list the objects (molecules, surfaces, pseudocentres, etc.) that are currently loaded.
- Each cavity opened will have a top-level branch in the main tree.
- Not all levels and branches of the tree need be displayed: clicking on any [-] icon will hide the details of the tree below that point, whereas clicking on a [+] icon will show more details.
- The display of individual objects can be controlled by clicking on the appropriate tick boxes adjacent to those objects.
- The surface and pseudocentre display settings made in the *Cavity Controls* window will automatically set the appropriate pseudocentre tick boxes in the *Graphics Object Explorer*. ***Note:*** these will be set in the individual tic boxes for each pseudocentre. A tick appears in the parent tick box for a pharmacophore type if any of the underlying pseudocentres are active. The tick box will be greyed unless all underlying pseudocntres are active.



3.3.2 Controlling which Molecules and Residues are Displayed

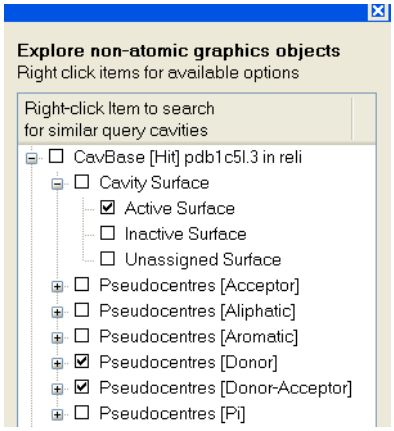
- The *Protein Explorer* window usually found to the top left of the hermes display can be used to control the display of individual ligand molecules, waters and proteins. The display of chains within proteins and individual amino acids residues can also be controlled (please refer to the Hermes documentation for further information).

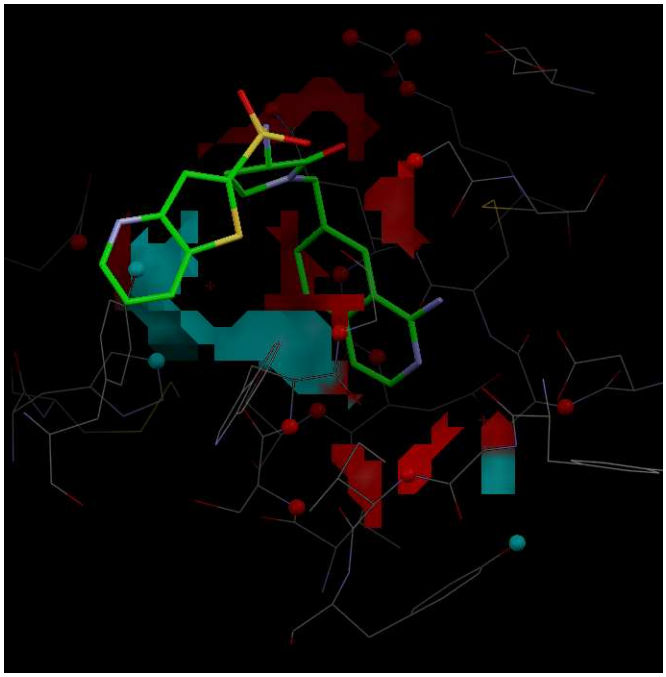
3.3.3 Controlling which Pseudocentres are Displayed

- The tick boxes in the *Graphics Object Explorer* window (bottom left of the Hermes window) can be used to switch on or off the display of particular types of pseudocentres (see Pseudocentres Section 1.4, page 98) (you may need to click on the [+] icon next to *P-Centres* to list the separate types).
- If the viewer is showing both the query cavity and a hit cavity from a similarity search, the display of pseudocentres may be further restricted to convey information about which pseudocentres in the query matched which in the hit, and which query pseudocentres were unmatched (see Displaying Matching and Non-matching Pseudocentres Section 3.2.3, page 103).

3.3.4 Displaying Parts of Surfaces

- First, some background information: by projecting each pseudocentre onto the cavity surface, the surface can be partitioned into patches, each patch associated with its closest pseudocentre (see Pseudocentres Section 1.4, page 98). There may be small gaps between these patches, i.e. some points on the surface may not be associated with any pseudocentre.
- The display of individual cavity surfaces is coupled to the display of pseudocentres, as explained below.
- If the two tick boxes *Inactive Surface* and *Unassigned Surface* in the *Graphics Object Explorer* (see below)) are turned off, only those parts of the surface that correspond to currently-displayed pseudocentres will be displayed. For instance the settings below turn on only the *Donor* and *Donor-Acceptor* pseudocentres. The surface can be turned off and on via the *Active Surface* tick box.





- If the *Inactive Surface* box is switched on, those parts of the surface corresponding to undisplayed pseudocentres will be shown (in green):
- If the *Unassigned Surface* box is switched on, those parts of the surface (if any) that are not associated with any pseudocentre will be displayed (in mauve):

3.3.5 Controlling Colour Schemes

- By default, surfaces are coloured by pseudocentre (i.e. each point on the surface is assigned the same colour as its nearest pseudocentre).
- It is possible to change the colour of a pseudocentre type by right-clicking the relevant type in the *Graphics Object Explorer*, and selecting an appropriate colour from the pull-down menu.

3.4 Tips: Useful Cavity-Viewer Settings

Two overlaid cavities present a complex visual image. When assessing how well a hit cavity matches with the query, the following strategies may help:

- Start by switching the surfaces off and displaying the matched pseudocentres only (to do this toggle on *Active Surface* and then toggle it off again. Do for both cavities). This should give you an immediate impression of how many of the query pseudocentres were matched, and how closely.
- Switch the surfaces back on but undisplay all but one type of pseudocentre, e.g. *Donor*. This will hide all parts of the surfaces except those corresponding to the pseudocentre type that you have left switched on, which makes it much easier to see how well the query and hit match. Obviously, this will need to be repeated for the other pseudocentre types. Since the aromatic, aliphatic and pi-types are chemically very similar, it is reasonable to inspect these parts of the surfaces together.
- To see how well a ligand fits into a cavity, it is useful to display the ligand in space filling mode together with all or part of the cavity surface.
- By default, carbon atoms in the query are shown in grey, those in the hit in green. Ligands are shown in stick mode, protein chains in wireframe. Different but related colour schemes are used for query and hit pseudocentres, e.g. blue for query donors and cyan for hit donors, red for query acceptors and pink for hit acceptors. Aliphatic, aromatic and pi pseudocentres are assigned the same colour by default since they are chemically similar. Query metal pseudocentres are coloured orange while hit metal pseudocentres are coloured yellow.

4 Cavity Similarity Searching

4.1 Overview of Cavity Similarity Searching

Similarity searching allows you to find cavities or parts of cavities in the database(s) that match a query cavity or sub-pocket. The query cavity itself must be taken from either the CavBase database that is supplied as part of this release (cavities from the PDB) or from an in-house cavity database that you have created. The steps in similarity searching are:

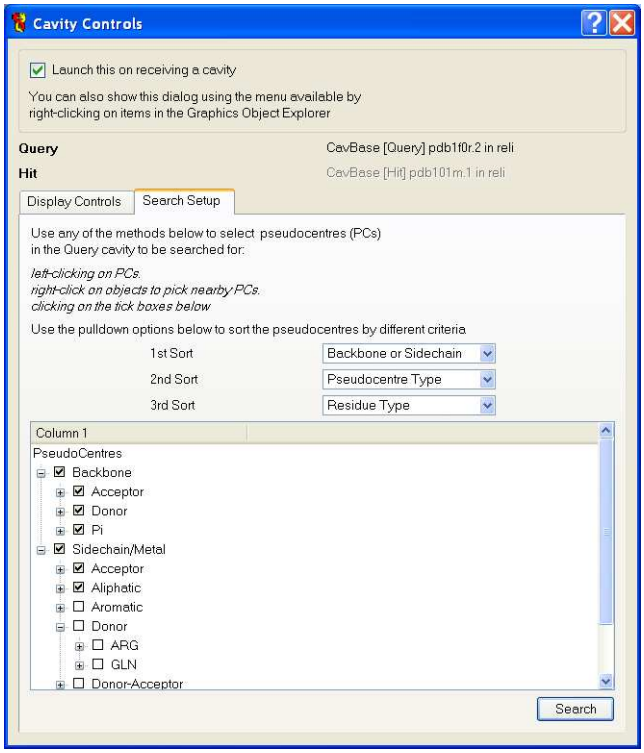
- Load the query cavity into Hermes (see Loading the Query Cavity Section 4.2, page 108).
- Select which of the pseudocentres you want to search for (see Selecting the Pseudocentres to be Searched For Section 4.3, page 108).
- Select other search options, e.g. how many hits to keep (see Setting Search Options and Starting the Search Section 4.4, page 110).
- Run the search.

4.2 Loading the Query Cavity

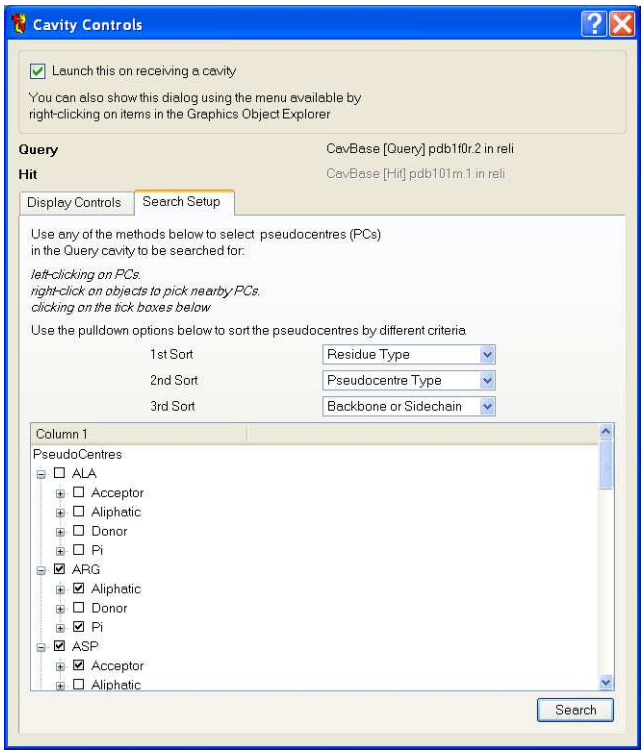
- Load the query cavity into the viewer by clicking on an appropriate link in a browser page (see Accessing Cavity Information for Relibase+ Database Entries Section 2, page 99).
- Move to the *Search Setup* pane of the viewer using the tab at the top-right of the *Cavity Controls* window:

4.3 Selecting the Pseudocentres to be Searched For

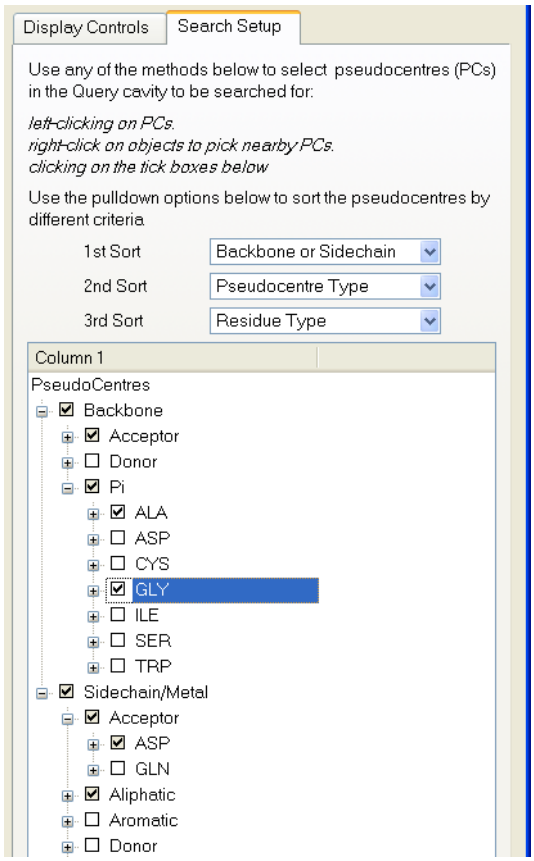
- When doing a cavity similarity search, it is not necessary to include all of the query-cavity pseudocentres in the search. For example, you can select just the pseudocentres in a sub-pocket; during the similarity search, all the other query-cavity pseudocentres will be completely ignored, i.e. the software will just try to find a good match for those in the sub-pocket. This is useful, for example, if you just want to find a good match for the immediate environment of a ligand that only occupies part of the query cavity.
- The selection of which query-cavity pseudocentres are to be included in the search is done by using the *Search Setup* pane of the *Cavity Controls* window. Unselected pseudocentres are displayed as translucent, picked pseudocentres are displayed as solid. The surface patch corresponding to each picked pseudocentre will be displayed. Pick the pseudocentres you want included in the search in any of the following ways:
  - Click on a pseudocentre with the left-hand mouse button to toggle its selection state.
  - To select all pseudocentres within a given distance of a pseudocentre, click on the pseudocentre with the right-hand mouse button and pick **Select to range** in the pull-down menu. Then select an appropriate distance.
  - To select all pseudocentres within range of an atom that has not a pseudocentre superimposed, or a bond; right-click on the atom or bond and then use the **Select Pseudocentres within range of this ligand...** option. You will be asked to enter a radius in the dialog window that appears.
  - The box at the base of the *Search Setup* pane allows you to select pseudocentres of particular types. The pull-down options above the white box control how the pseudocentres are listed, e.g. the settings below will cause pseudocentres to be listed, firstly by whether they arise from backbone or side-chain, secondly, by the pseudocente type, and thirdly, by the type of residue that they belong to:



- The tick boxes may then be used to select particular types of pseudocentres, e.g. the following settings indicate that donor pseudocentres in Asn and Asp residues are selected:



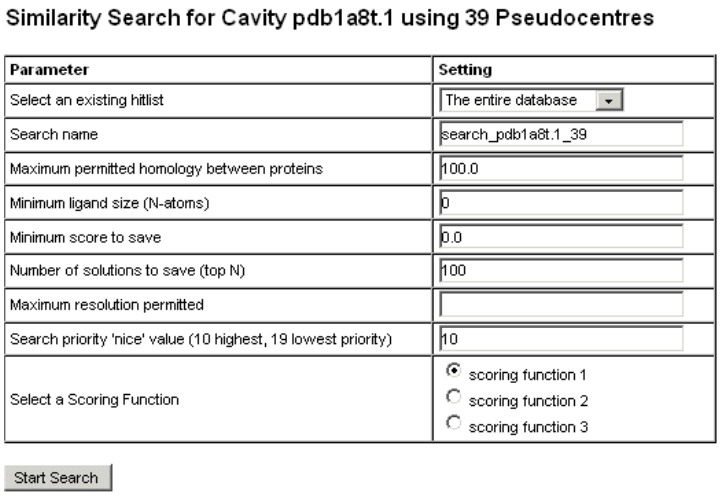
- Because pseudocentres can also be selected and deselected in other ways, e.g. by clicking on them, it is possible to generate a state in which some pseudocentres of a given type are selected and some are deselected. This is indicated by a tick box on a grey background, e.g.



- Once you have selected the pseudocentres you want to search for, hit the **Search** button at the bottom right of Hermes. This will open a browser page where you can set some other search options and then start the search.

4.4 Setting Search Options and Starting the Search

- Once you have specified the search query, i.e. hit the **Search** button in the *Search Setup* pane of the *Cavity Controls* pane (see Selecting the Pseudocentres to be Searched For Section 4.3, page 108), you will be taken to a browser page looking something like this:



- This page can be used to set a variety of search options, as follows:
  - Use the *Select a cavity hitlist* pull-down menu to specify whether you want to search the whole database (pick *The entire database* from the pull-down menu) or a subset. In order to search a subset, you must have previously created a hitlist defining that subset (see Searching a Subset of the Database Section 4.4.1, page 112); if you have done this, simply pick the relevant hitlist from the pull-down menu.
  - Type in a name for the search.
  - Specify the maximum permitted homology, as a percentage. This can be used to increase the novelty of the search results by rejecting hit cavities from proteins that are similar in sequence to the query-cavity protein. Such hits are often trivial and can be found more quickly by sequence-based similarity search methods.
  - Specify whether hit cavities are to be rejected if they are not occupied by ligands of at least *N* atoms.
  - Type in the minimum permitted score; hit cavities will only be kept if their similarity with the query exceeds this value. Unfortunately, the different scoring functions are on different scales, so you will need to get some experience before you can use this option effectively.
  - Specify the maximum number of hits that you want, e.g. the top 50.
  - Specify the maximum allowed resolution.
  - Specify the *Search priority* value (10 highest, 19 lowest priority). This option effectively allows you to determine how important searches are so that they either run quickly (top priority) or run in the background without interfering with other tasks (low priority).
  - Select a scoring function (see Similarity Searching and Scoring Section 1.5, page 99). At present, the performance of these functions has not been fully characterised, so you may need to experiment to find which one works best for any particular query. All of them seem reasonably reliable; at CCDC, we usually use scoring function 3.

- Hit **Start Search** to begin the search. Searches may take many hours; it is safe to close down your Relibase+ session while a search is running (i.e. it will not stop the search and you can collect the results later from a new session).

Related topics:

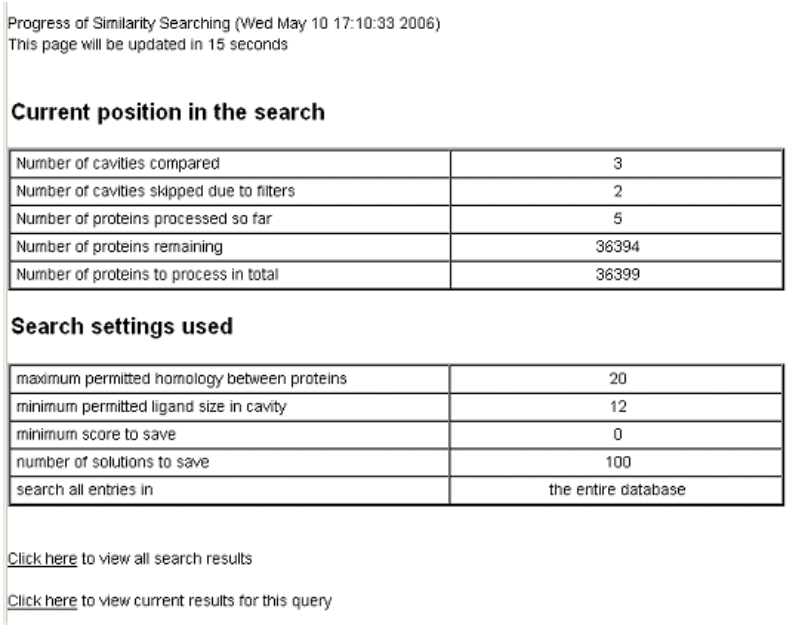
- Searching on a Subset of the Database; Cavity Hitlists (see Searching a Subset of the Database Section 4.4.1, page 112)
- Monitoring Search Progress; Aborting Searches (see Monitoring Search Progress; Aborting Searches Section 4.5, page 112)
- Browsing the Results of a Search (see Browsing the Results of a Search Section 4.6, page 113)

4.4.1 Searching a Subset of the Database

- Cavity similarity searches can take a very long time, so if you know in advance that you only want hits from certain types of proteins, e.g. kinases, you should confine the search to a subset of just those entries.
- To set up a subset, you need to create a hitlist. Subsets or hitlists that can be searched are protein or ligand hitlists, or hitlists that have been converted to cavity hitlists.
  Note: it is no longer essential to convert a protein or ligand hitlist to a cavity hitlist prior to using the hitlist as a cavity search subset.
- Start by performing a Relibase+ search to find the entries you want, e.g. a text search for the keyword *kinase*. Save the results in a protein hitlist (if you are searching on a protein property) or a ligand hitlist (if you are searching on a ligand property).
- Select the relevant protein or ligand hitlist from the *Select an existing hitlist* pull-down menu. It will then be used as a subset for the cavity similarity search.

4.5 Monitoring Search Progress; Aborting Searches

- Once you have started a cavity similarity search, you will see a display something like this:



Progress of Similarity Searching (Wed May 10 17:10:33 2006)
This page will be updated in 15 seconds

Current position in the search

Number of cavities compared	3
Number of cavities skipped due to filters	2
Number of proteins processed so far	5
Number of proteins remaining	36394
Number of proteins to process in total	36399

Search settings used

maximum permitted homology between proteins	20
minimum permitted ligand size in cavity	12
minimum score to save	0
number of solutions to save	100
search all entries in	the entire database

Click here to view all search results

Click here to view current results for this query

- The display will be updated every 15 seconds so that you can monitor the progress of the search. Note that the number of cavities to be searched will invariably exceed the number of proteins, since many proteins contain more than one cavity.
- To view the results so far, click on the hyperlink *Click here to view current results for this query.* This will take you to a page listing the hits in descending order of similarity (see Browsing the Results of a Search Section 4.6, page 113). Clicking on any hit will load it and the query into Hermes.
- Cavity similarity searches can take many hours to run. If you want to stop a search before it has finished, click on *Finish this query now.* Any hits already found will be kept.

4.6 Browsing the Results of a Search

- If you start a search and then keep your Relibase+ session open at the search progress page (see Monitoring Search Progress; Aborting Searches Section 4.5, page 112), you will, on completion, automatically be shown a table summarising the search results. If you exit Relibase+ while the search is in progress, you can use the search management tool to access the same table (see Managing Search Results Section 4.7, page 116).
- A summary of the search settings is given at the top of the *Cavity Comparison Search Result* window, followed by a table of search results. The results table will look something like this:

Search Results 1-15 of 400

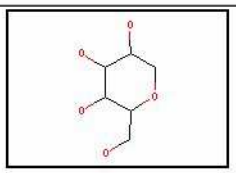
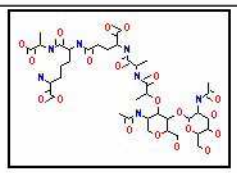
Select	Cavity	Score	Normalised Score	Matched Centres	RMS	Protein Homology	Cavity Homology	Header	Title
☑	pdb1bxg.1	5806.0	97.5	35	0.187	100.0	100.0	HYDROLASE	ACID PROTEINASE (PENICILLOPEPSIN) COMPLEX WITH PHOSPHONATE INHIBITOR.
☐	pdb1ppl.1	5141.7	86.4	33	0.569	100.0	78.0	HYDROLASE(ACID PROTEINASE)	CRYSTALLOGRAPHIC ANALYSIS OF TRANSITION-STATE MIMICS BOUND TO PENICILLOPEPSIN: PHOSPHORUS-CONTAINING PEPTIDE ANALOGUES
☐	pdb2web.1	5054.7	84.9	32	0.565	100.0	90.2	HYDROLASE	ACID PROTEINASE (PENICILLOPEPSIN) (E.C.3.4.23.20) COMPLEX WITH PHOSPHONATE INHIBITOR: METHYL(2S)-[1-(((N-FORMYL)-L VALYL)AMINO-2-(2-NAPHTHYL)ETHYL)HYDROXYPHOSPHINYLOXY]-3 PHENYLPROPANOATE, SODIUM SALT
☐	pdb1ppm.1	4963.5	83.4	33	0.528	100.0	90.2	HYDROLASE(ACID PROTEINASE)	CRYSTALLOGRAPHIC ANALYSIS OF TRANSITION-STATE MIMICS BOUND TO PENICILLOPEPSIN: PHOSPHORUS-CONTAINING PEPTIDE ANALOGUES
☐	pdb1ppk.1	4876.2	81.9	31	0.453	100.0	92.7	HYDROLASE(ACID PROTEINASE)	CRYSTALLOGRAPHIC ANALYSIS OF TRANSITION STATE MIMICS BOUND TO PENICILLOPEPSIN: PHOSPHOROUS-CONTAINING PEPTIDE ANALOGUES

- Each row of the table relates to a hit cavity found by the search. By default, the hits will be sorted in descending order of similarity score. However, you can click on any of the columns to sort the table on that column.
- The table gives the following information:
  - *Cavity*: identifier of the hit cavity.
  - *Score*: the similarity score for the cavity from the chosen scoring function.
  - *Normalised Score*: the similarity comparison score, normalised with respect to the query cavity, given as a percentage.
  - *Matched centres*: the number of matching pseudocentres in the superimposed cavities.
  - *RMS*: the root mean square deviation resulting from superposition of the matching pseudocentres.
  - *Protein Homology*: the percentage sequence identity of complete protein chains defining the query and hit cavity binding site. The identity is calculated for all pairs of protein chains in the query and hit cavity and the protein homology is the highest of these values.
    Note: a subtlety in the method used for the homology calculation means that only residues for which there are coordinates in the entry are included in the calculation, i.e. the calculation may be based on fewer residues than for homology values obtained via another method (either a protein sequence similarity search or from a similar binding site search). This may give different homology values depending on the calculation method used.
  - *Cavity Homology*: the percentage similarity between all the protein chains defining the query cavity and the hit cavity. The complete protein chains are aligned and the identity is calculated using only those residues in the cavity binding site.
  - *Header*: header information taken from the PDB file.
  - *Title*: the title record for the PDB file.

- A user-specified list of cavities can be selected for display or for download using the tick boxes in the first column of the table. Individual cavities can be selected by activating the corresponding tick box; cavities can be selected or deselected globally using the **Select** button at the top of the column. After the desired cavities have been selected, use the **Superpose Selected Cavity Binding Sites** button at the bottom of the table to expose the following options:
  - **Display Superposed Cavity Binding Sites in Hermes**: click on this link to load the selected cavities into Hermes.
  - **Download Superposed Cavity Binding Site in Mol2 Format**: click on this link to download the selected cavities to a mol2 format file.
- If you change your mind and update your cavity selections, use the **Superpose Selected Cavity Binding Sites** button to refresh the contents of the files to be displayed in Hermes or to be downloaded.
- To see more information about a hit, click on the relevant cavity identifier in column 1. This will load both the hit and the query cavity into the 3D cavity viewer (see Displaying and Comparing Cavities Section 3, page 101) and will also display two summary tables, e.g.

Comparison of cavities pdb1xie.3 and pdb148l.1

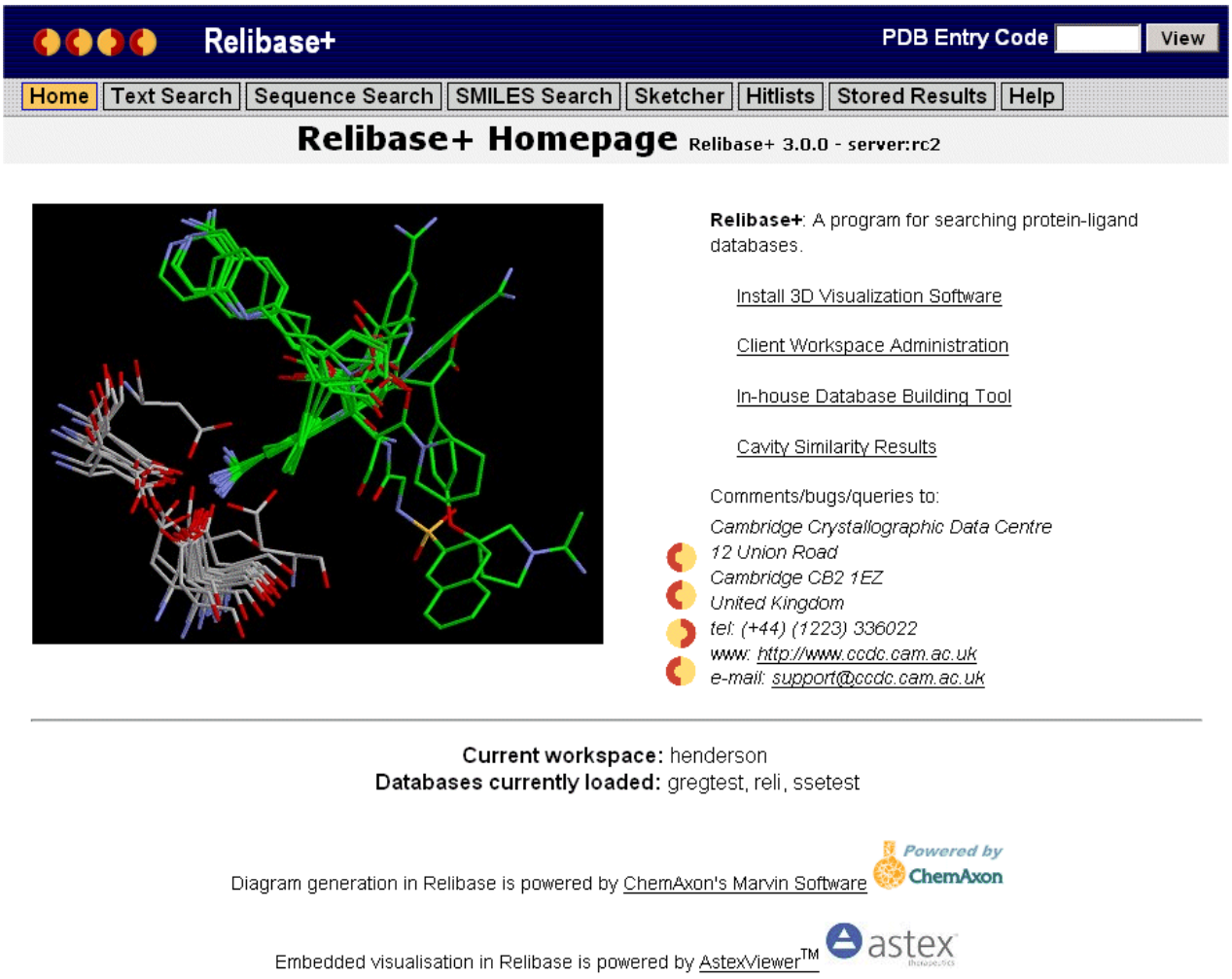
Pseudo-centres (match)	6
RMS (match)	0.79
Pseudo-centres (clique)	6
RMS (clique)	0.79
Score	5.18
Protein Homology	12.10
To reload comparison	click here

	Query cavity pdb1xie.3		**Similar cavity pdb148l.1**
Volume ($Å^3$)	2701.50		965.12
Pseudo-centres (total)	190		62
PDB	1xie		148l
Header	ISOMERASE(INTRAMOLECULAR OXIDOREDUCTSE)		HYDROLASE(O-GLYCOSYL)
Title			
Compound	D-XYLOSE ISOMERASE (E.C.5.3.1.5) COMPLEXED WITH 1,5-DIANHYDROSORBITOL (PH 7.4)		LYSOZYME (E.C.3.2.1.17) MUTANT WITH THR 26 REPLACED BY GLU CYS 54 REPLACED BY THR, AND CYS 97 REPLACED BY ALA (T26E,C54T,C97A) COMPLEXED WITH SUBSTRATE CLEAVED FROM CELL WALL OF ESCHERICHIA COLI
Ligands			
Cavity object	go to cavity information page		go to cavity information page

- The search program uses a clique-detection algorithm to produce a preliminary mapping of query- and hit-cavity pseudocentres; the entries *Pseudo-centres (clique)* and *RMS (clique)* refer to the number of pseudocentre pairs in this match and their RMS deviation, respectively. Additional pseudocentres may be added to this preliminary mapping, or pseudocentres may be dropped from it, and *Pseudo-centres (match)* and *RMS (match)* give the number of pairs and RMS deviation for this final mapping. All the other information in the tables should be self-evident.
- Clicking on a protein identifier or a ligand diagram will take you to the relevant Protein or Ligand Information page.

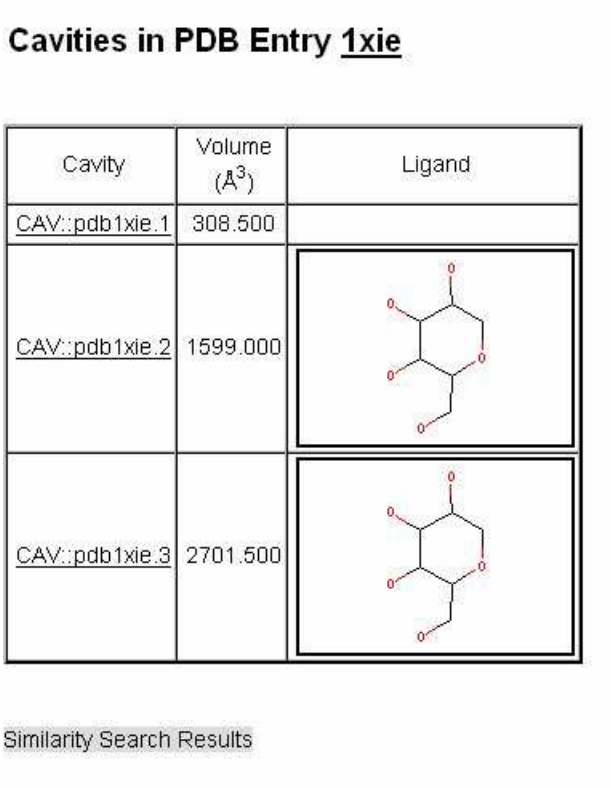
4.7 Managing Search Results

- To see a list of all the cavity similarity searches you have run, click on the *Cavity Similarity Results* hyperlink on the Relibase+ home page:



- A link to cavity similarity search results is also available via the **Stored Results** tab.

- Hyperlinks to the cavity similarity search results list can be found on any cavity information page (see Accessing Cavity Information for Relibase+ Database Entries Section 2, page 99), e.g.



- An example list of cavity similarity searches is:

Comparisons for user 'robertson'

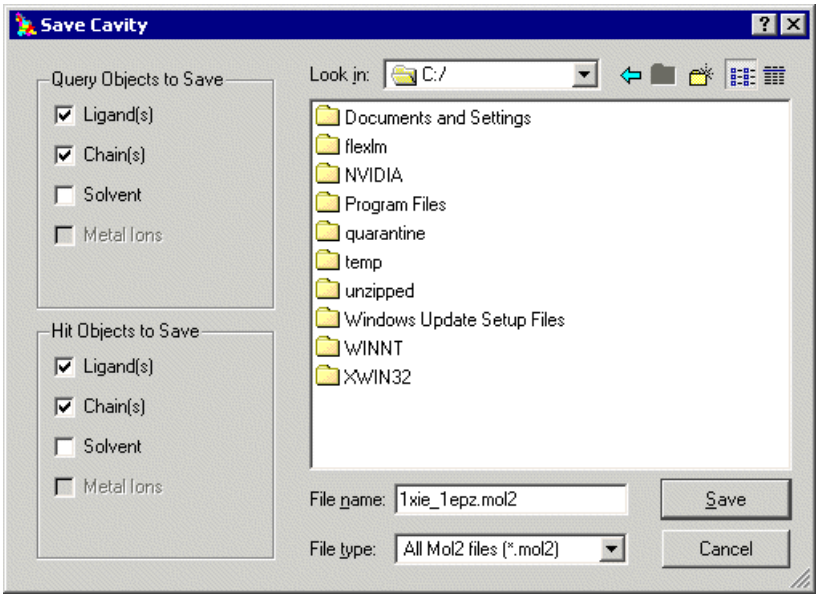
Query Name	Query Cavity	Number of Comparisons	Search Date	Delete
search_pdb1jus.5_49_2	pdb1jus.5	100	May 10 17:15	☐
search_pdb1jus.5_49	pdb1jus.5	100	May 10 16:51	☐
tutorial_1jus	pdb1jus.5	100	May 9 15:52	☐
search_pdb1dm8.9_6	pdb1dm8.9	29	Apr 26 15:40	☐

- Clicking on any item in the *Query Name* column will show the hits from that particular search, which you can then browse and view in 3D (see Browsing the Results of a Search Section 4.6, page 113).
- Clicking on an item in the *Query Cavity* column will take you to the cavity information page for that cavity (see Accessing Cavity Information for Relibase+ Database Entries Section 2, page 99) and will load the cavity into the 3D cavity viewer (see Displaying and Comparing Cavities Section 3, page 101).

- Ticking an entry in the final column and then clicking on the **Delete** button will permanently remove the results of that search from your workspace. *(Caution: there is no undo facility or request for confirmation.)*

5 Saving Cavities to File

- The molecular components (ligands, chains, solvent molecules) of a cavity, or of an overlaid pair of query and hit cavities, can be saved in .mol2 format from Hermes. Load the cavity (see Accessing Cavity Information for Relibase+ Database Entries Section 2, page 99) or the query-hit pair (see Browsing the Results of a Search Section 4.6, page 113), then select the top-level menu option **File**, followed by **Save Cavity** in the pull-down menu.
- Use the tick-boxes in the resulting dialogue to specify which components you want to write out, specify an output file name, then hit **Save**, e.g.



6 Building In-House Cavity Databases

Cavity information is generated for proprietary structures by default when the structures are processed to an in-house database. These data are then searchable when using the Cavity Information Module. Further information on producing proprietary databases is provided elsewhere (see CHAPTER 6: CREATING IN-HOUSE DATABASES Section , page 157).

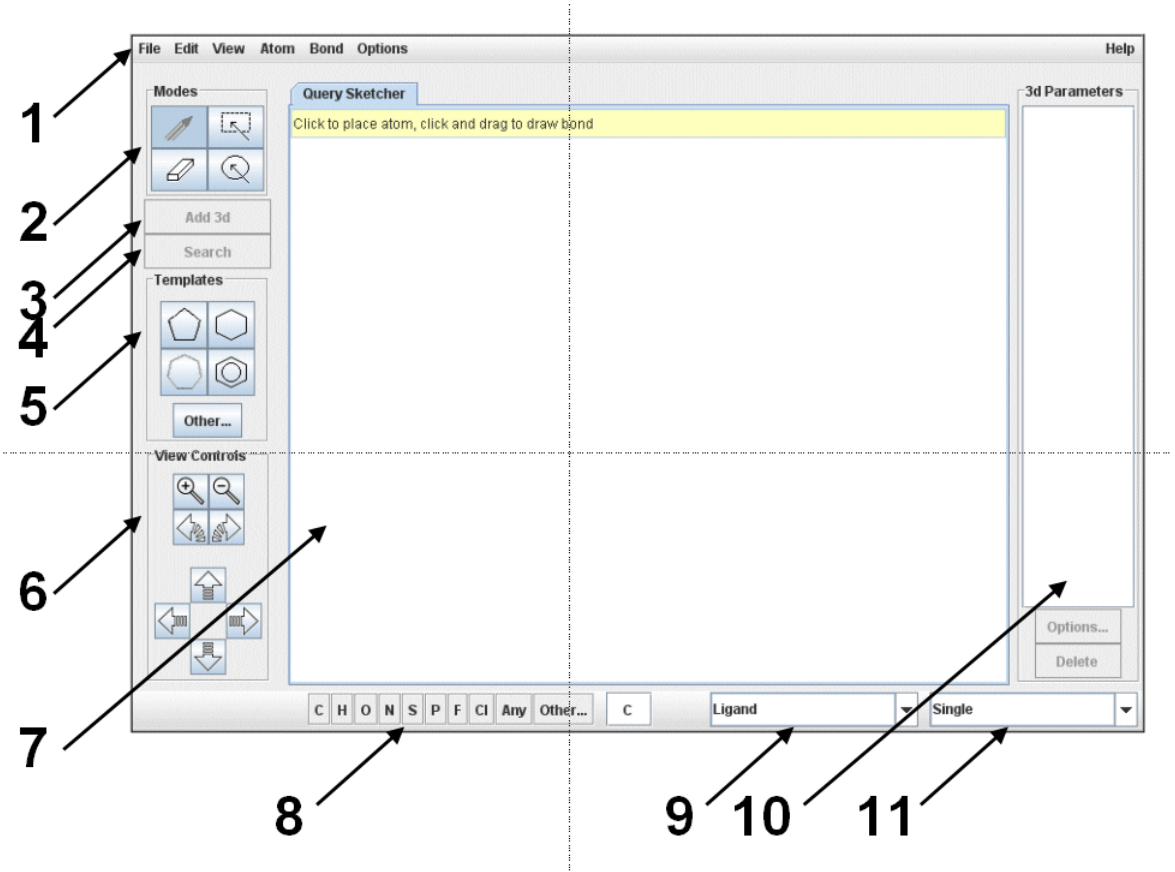
CHAPTER 5: USING THE RELIBASE SKETCHER

1 **Sketcher Basics** (see page 119)
2 **Fundamentals of Drawing** (see page 122)
3 **Drawing and Fusing Rings** (see page 125)
4 **Atom Properties** (see page 127)
5 **Bond Properties** (see page 131)
6 **Using Substructure Templates** (see page 133)
7 **Substructure Display Conventions** (see page 134)
8 **Moving, Scaling, Rotating and Duplicating Substructures** (see page 135)
9 **Reading, Saving and Deleting Queries** (see page 137)
10 **Geometric Objects** (see page 138)
11 **Geometric Parameters** (see page 139)
12 **Applying Constraints** (see page 144)

1 Sketcher Basics

1.1 Layout of the 2D/3D Drawing Window

Open the substructure drawing window by clicking on the **Sketcher** button in the Relibase+ menubar.



1. Top-level menu..
2. Mode buttons - responses to mouse clicks in the drawing area will depend on which mode is active (see Modes in the Drawing Window Section 1.2, page 120).
3. Buttons to set up geometrical parameters and constraints (see Geometric Parameters Section 11, page 139).
4. Button for starting searches (see Running a Search Section 6.8, page 74).
5. Menu for selecting templates (molecular building blocks to aid drawing) (see Using Substructure Templates Section 6, page 133).
6. View controls - buttons to translate, rotate and re-size the display.
7. Drawing area (see Sketcher Basics Section 1, page 119).
8. Area for changing the current element type (see Changing the Current Element Type Section 4.1, page 127).
9. Menu for selecting molecule type (water, ligand or protein) (see Setting Molecule Types Section 1.3, page 120).
10. Area for listing, displaying and editing 3D and nonbonded contact parameters (see Geometric Parameters Section 11, page 139).
11. Area for changing the current bond type (see Changing the Current Bond Type Section 5.1, page 131).

1.2 Modes in the Drawing Window

The four *Mode* buttons on the top left-hand side of the *Sketcher* window are mode buttons which affect what happens when the mouse is used in the drawing area.

- **Draw**: click on this button when you want to draw a substructure.
- **Select:** click on this button when you want to perform editing tasks such as moving, or resizing substructures, or selecting atoms or bonds.
- **Lasso:** as select but the selection area becomes a user-defined shape rather than a rectangular panel.
- **Delete**: click on this button when you want to delete atoms or bonds.

1.3 Setting Molecule Types

- When using the sketcher you must ensure that the molecule type is set correctly for each substructure that is drawn**.**
- There are three searchable molecule types in Relibase+: **Protein**, **Ligand**, and **Water** (nucleic acids are not searchable).
- *Note:* In Relibase+, all moieties which are neither protein nor nucleic acid in a structure are considered to be ligands. Hence metal ions, anions, solvate molecules (except water), cofactors and inhibitors are all regarded as ligands. The molecule type **Ligand** must therefore be used when searching for these moieties.
- Protein substructures are displayed in blue, ligand substructures in black and water atoms in pink

in the sketcher window.

- The current molecule type may be changed by clicking on the button at the bottom of the *Draw* window and selecting from the resulting pull-down menu.

- It is also possible to allow substructures to be either **Protein or Ligand, Protein or Water,** or **Ligand or Water** molecule types. To allow a substructure to be either **Protein or Ligand or Water** select the molecule type **Any**. Substructures of these mixed types are drawn in grey.
- The selected molecule type will determine the type of any new atom created when drawing (in **Draw** mode)
- There are a number of ways to change the molecule type of an existing substructure, including:
  - In **Select** mode, Select the substructure that you wish to change (see Selecting Atoms Section 2.7, page 123), then change the current molecule type using the pull-down menu at the bottom of the *Draw* window.
  - In any mode, right click on an atom or bond within the substructure and from the resulting menu select either **Place Fragment in Ligand** or **Place Fragment in Protein**.
  - To set the molecule type of all atoms, in any mode, pick **Atoms** from the top-level menu, and select either **Place All Atoms in Ligand** or **Place All Atoms in Protein** from the resulting pull-down menu.

2 Fundamentals of Drawing

2.1 Drawing a Bond

- Ensure you are in **Draw** mode (see Modes in the Drawing Window Section 1.2, page 120).
- Ensure the molecule type is set appropriately, i.e. protein, ligand, or water (see Setting Molecule Types Section 1.3, page 120).
- Move the cursor into the white area of the drawing window.
- Press down the left-hand mouse button, move the cursor *while keeping the mouse button depressed*, and then release the button.
- This draws a bond, using the appropriate element and bond types (see Changing the Current Element Type Section 4.1, page 121 or Changing the Current Bond Type Section 5.1 page 125)
- To draw bonds of fixed length, select **bond length** from the **Options** menu and tick the **Fixed** box. The length **Standard/Half/Double** can be selected using the radio buttons.
- To draw bonds which are fitted to a grid, select **Grid** in the top-level **View** option and click on the specific grid required (**Horizontal triangles**, **Vertical triangles** or **Square**). Select **No Grid** to remove. The **GridSize** slider alters the Grid Size. To change its orientation, type the rotation angle into the box.

2.2 Drawing an Isolated Atom

- Ensure you are in **Draw** mode (see Modes in the Drawing Window Section 1.2, page 120).
- Ensure the molecule type is set appropriately, i.e. protein, ligand, or water (see Setting Molecule Types Section 1.3, page 120).
- Move the cursor into the white area of the drawing window.
- Click the left-hand mouse button, and release it again without moving the mouse.

2.3 Drawing a Bond from an Existing Atom

- Ensure you are in **Draw** mode (see Modes in the Drawing Window Section 1.2, page 120).
- Ensure the molecule type is set to that of the existing atom, i.e. protein, ligand, or water (see Setting Molecule Types Section 1.3, page 120).
- Move the cursor onto the atom.
- Press down the left-hand mouse button.
- Move the cursor while keeping the mouse button depressed, then release the button.

2.4 Drawing a Bond to an Existing Atom

- Ensure you are in **Draw** mode (see Modes in the Drawing Window Section 1.2, page 120).
- Ensure the molecule type is set to that of the existing atom, i.e. protein, ligand, or water (see Setting Molecule Types Section 1.3, page 120).

- Move the cursor into the white area of the drawing window.
- Press down the left-hand mouse button.
- Move the cursor onto the desired atom (the bond *locks* onto the atom) while keeping the mouse button depressed, then release the button.

2.5 Drawing a Bond between Two Existing Atoms

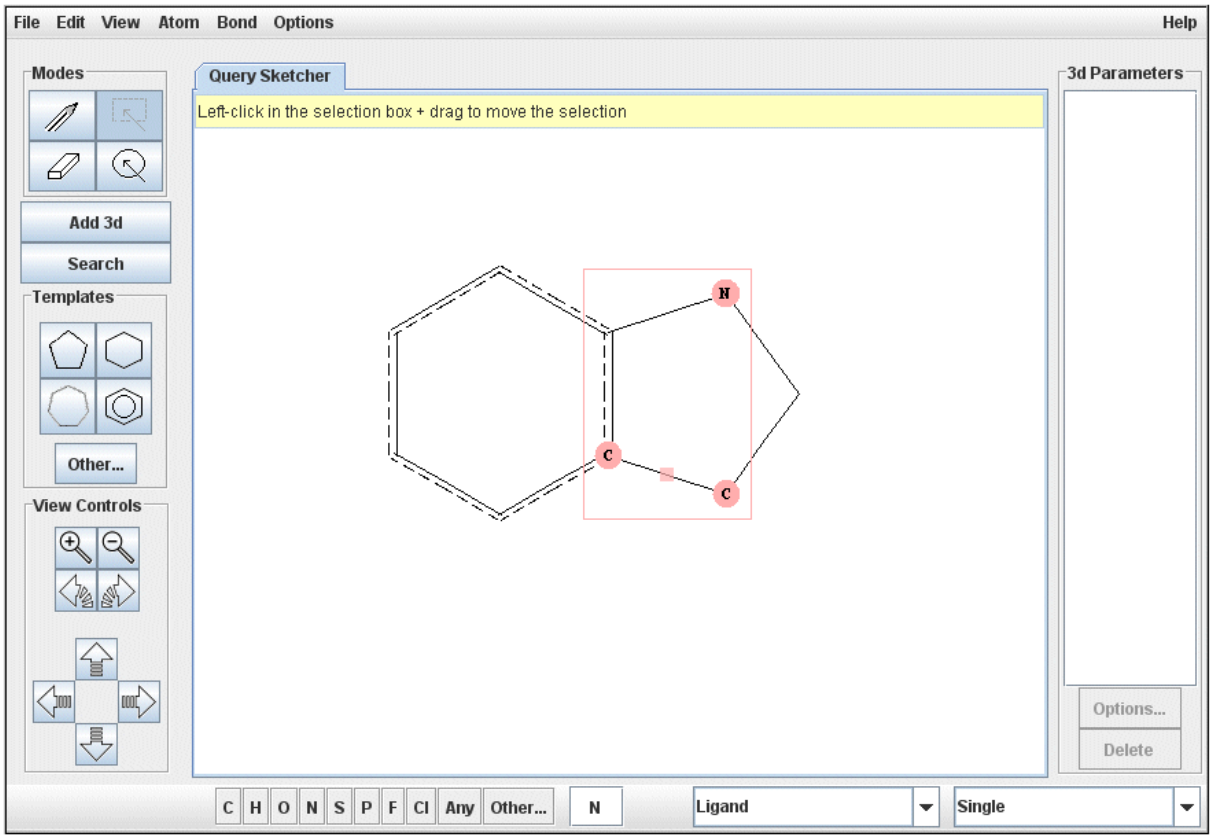
- Ensure you are in **Draw** mode (see Modes in the Drawing Window Section 1.2, page 120).
- Ensure the molecule type is set to that of the existing atoms, i.e. protein, ligand, or water (see Setting Molecule Types Section 1.3, page 120). It is not possible to add bonds between different molecule types.
- Move the cursor onto the first atom.
- Press down the left-hand mouse button.
- Move the cursor onto the second atom (the bond *locks* onto the atom) while keeping the button depressed, then release the button.

2.6 Undoing Mistakes when Drawing Substructures

- Select **Edit** in the top-level menu and **Undo** in the resulting pull-down menu to undo the last action performed.
- If necessary, **Edit... Undo** may be used several times in a row to undo a sequence of actions, one by one.

2.7 Selecting Atoms

- Selection of atoms is useful for setting the molecule type of atoms (water, ligand or protein), deleting atoms, moving substructures around the drawing area and for superposition of hits. Selected atoms are shown in pink:



Atoms and bonds may be selected in several ways:

- In **Select** mode, an individual atom can be selected or deselected by clicking on it with the left-hand mouse button.
- In **Select** mode, a series of atoms or bonds can be selected by clicking on each in turn while keeping the Shift key pressed down.
- In **Select** mode, a group of atoms and bonds can be selected by clicking with the left-hand mouse button on a blank point in the white area and moving the cursor while keeping the mouse button pressed down. Everything enclosed in the resulting rectangular box gets selected when the mouse button is released.
- Groups of atoms within a non-rectangular shape can also be selected in **Lasso** mode.
- In any mode, everything can be selected by hitting **Edit** in the top-level menu and **Select All** in the resulting pull-down menu (or by using the CTRL-A shortcut).
- In any mode, the current selection can be reversed by hitting **Edit** in the top-level menu and **Invert Selection** in the resulting pull-down menu (or by using the CTRL-I shortcut). Everything that was selected becomes unselected, and vice versa.
- In any mode, everything can be deselected by hitting **Edit** in the top-level menu and **Deselect All** in the resulting pull-down menu (keyboard shortcut - CTRL-SHIFT-A).
- In any mode, a bonded fragment may be selected by right-clicking on an atom or bond in the

fragment and selecting **Select Fragment**.

2.8 Deleting Atoms and Bonds

There are several methods, including:

- In **Delete** mode, click with the left-hand mouse button on the atom or bond to be deleted.
- In any mode, click with the right-hand mouse button on the atom or bond to be deleted and pick **Delete Atom** or **Delete Bond** from the resulting pull-down menu.
- Select the atoms and bonds that you wish to delete (see Selecting Atoms Section 2.7, page 123), then click with the right-hand mouse button on a blank point in the white area and pick **Delete Selected** from the resulting pull-down menu. (or by using the keyboard **Delete** shortcut).
- To delete all atoms and bonds, in any mode, move the cursor onto a blank point in the white area, click on the right-hand mouse button, and pick **Delete All** from the pulldown menu.

3 Drawing and Fusing Rings

3.1 Adding a Ring to a Blank Drawing Area

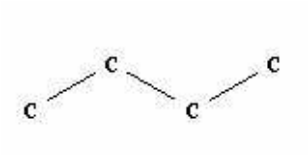
- Rings may be drawn manually but the easiest way is to use the pre-drawn rings to the left of the *Draw* window:

- If the desired ring is one of the four on display (see above), select it by clicking on its icon, move the cursor into the white area, then click with the left-hand mouse button.
- Click on the **Draw** button to stop drawing rings.
- Some complex ring systems are available by clicking on **Other...** in the *Templates* section to the left of the *Draw* window.

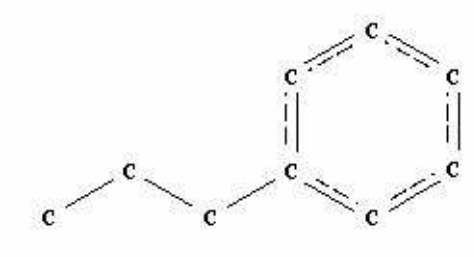
3.2 Adding a Ring to an Atom in an Existing Substructure

- Select the ring (see Adding a Ring to a Blank Drawing Area Section 3.1, page 125), then click on the desired atom in the existing substructure with the left-hand mouse button.

- For example, selecting a 6-membered aromatic ring and clicking on the terminal C atom in:

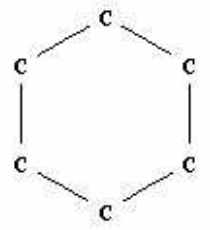


will create:

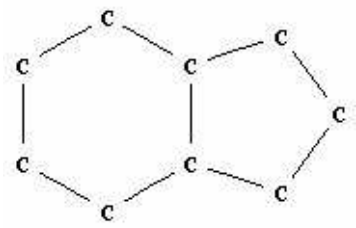


3.3 Fusing a New Ring to an Existing Ring

- Select the new ring (see Adding a Ring to a Blank Drawing Area Section 3.1, page 125), then click on the desired fusion bond in the existing ring.
- For example, selecting a 5-membered saturated ring and clicking on one of the C-C bonds in:



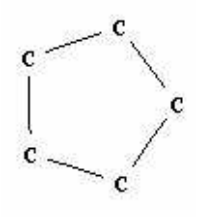
will create:



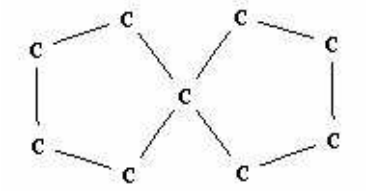
3.4 Creating a Spiro-Fusion

- Select the required ring (see Adding a Ring to a Blank Drawing Area Section 3.1, page 125), then click on the desired spiro atom in an existing ring.

- For example, selecting a 5-membered saturated ring and clicking on one of the C atoms in:



will create:



3.5 Fusing Rings by Moving One Ring onto Another

- It is possible to fuse two separate rings in the drawing area by selecting all the atoms in one ring (see Selecting Atoms Section 2.7, page 123) and moving it towards the other (see Moving Atoms Section 8.1, page 135).
- Spiro fusion is achieved by overlapping one atom in the moveable ring with one atom in the stationary ring (indicated by the overlapped atoms being highlighted). Fusion will occur when the mouse button is released.
- Bond fusion is achieved by overlapping two bonded atoms in the moveable ring with two bonded atoms in the stationary ring. It may be necessary to overlap one of the pairs and then rotate the moveable ring (by holding down the Control key) until the second pair overlap.

4 Atom Properties

4.1 Changing the Current Element Type

- The current element type determines the type of any new atom created when drawing (in **Draw** mode). The current setting is shown in the white box at the bottom of the *Draw* window.

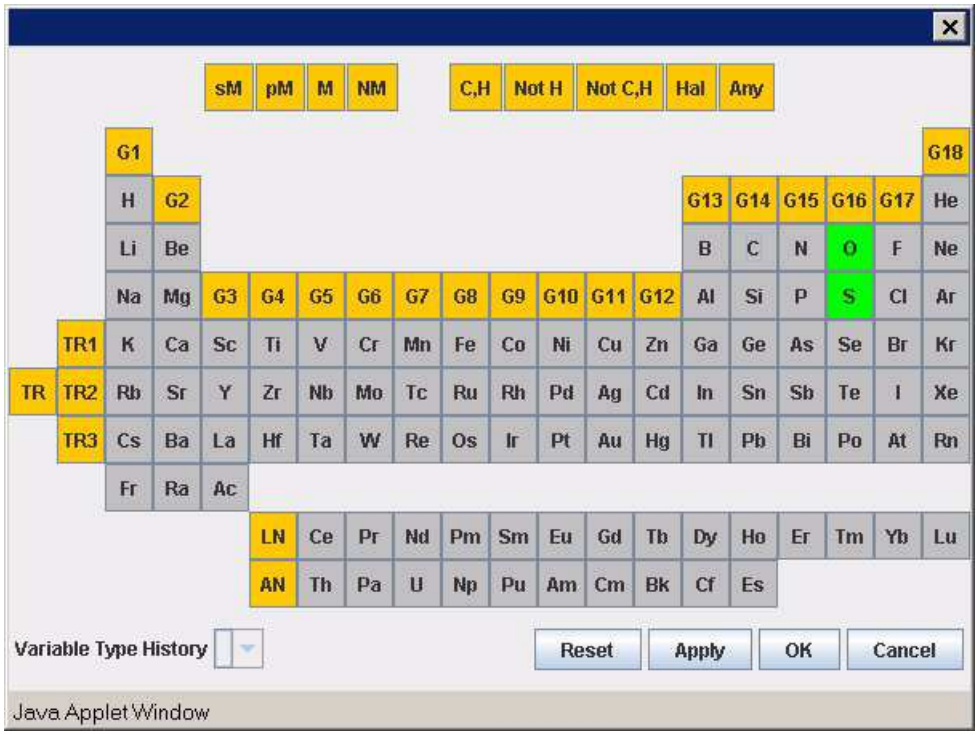
- The current element type may be changed by hitting any of the element symbols at the bottom of

the *Draw* window.

- *Any* means any atom (i.e. an atom of any element type), and is denoted by the symbol **Any** in the drawn substructure.
- **Other...** displays a pull-down menu. Choosing **Select from periodic table...** allows selection of any element in the Periodic Table.
- Right-click on the canvas, ensuring no atoms are selected, click on **Set Element Type** and follow the pull down menu to the appropriate atom type.

4.2 Setting Variable Element Types

- Atoms in a substructure may be variable, e.g. F *or* Cl *or* Br *or* I.
- *Any* means any atom (i.e. an atom of any element type), and is denoted by the symbol **Any** in the drawn substructure.
- The current element type (see Changing the Current Element Type Section 4.1, page 127) may be made variable by hitting the **Other...** button (at the bottom of the *Draw* window). This displays a menu from which common variable element types (e.g. **Any Metal**, Halogen etc.) can be selected. Alternatively, choosing **Select from periodic table...** from this menu opens up the Periodic Table:



- From here it is possible to create your own variable element type by clicking on the required elements, e.g. O and S. Pre-defined element groups can also be used by selecting the appropriate group symbols from the periodic table.

- Hit **Apply** to accept the currently selected element types for drawing, or **OK** to accept the currently selected element types and close the periodic table. The resulting variable element type is called *V1*, if it is the first variable type created, *V2*, if it is the second, etc.

4.3 Changing the Element Types of Existing Atoms

This can be done in several ways, including:

- In **Any** mode, click on the atom with the right-hand mouse button and select **Set element type...** from the resulting pull-down menu, then select the required element type. **Other...** allows selection of some pre-defined variable element types or selection from the periodic table (see Setting Variable Element Types Section 4.2, page 128).
- In **Draw** mode, change the current element type (see Changing the Current Element Type Section 4.1, page 127) and then click on the atom with the left-hand mouse button.
- In **Any** mode, click on **Atoms** in the top-level menu, select **Element** from the resulting pull-down menu and select the required element. **Other...** allows selection of some pre-defined variable element types or selection from the periodic table (see Changing the Current Element Type Section 4.1, page 127). The *Atom Property* pop-up appears, click on the atom or atoms to be changed with the left-hand mouse button and hit **Done**.
- Select atoms and either click on **Atoms** in the top level menu, select **Element** from the resulting pulldown menu and select the required element, or right-click on the canvas and select **Set element type**.

4.4 Addition of Hydrogen Atoms

- Hydrogen atoms may be drawn in the same way as any other type of atom or they may be defined implicitly. It is only possible to add implicit hydrogens to carbon atoms since the number of hydrogens on heteroatoms cannot be safely inferred.
  Note: if using protein and ligand templates, H atoms are already defined implicitly on the templates themselves (see Using Substructure Templates Section 6, page 133).
- To add hydrogen atoms implicitly to carbon enter **Draw** or **Select** mode, click on an atom with the right-hand mouse button, pick **Number of Hydrogens on Carbon** from the resulting pull-down menu, then select the number of hydrogens required from the second pull-down menu. **Unspecified** in this menu means that any number of hydrogens is allowed. Selecting **Generate** automatically adds the appropriate number of hydrogen atoms to the carbon atoms so as to satisfy the valency requirements; picking **Clear** removes them.
- Hydrogen atoms can also be added to any carbon atom already drawn on the sketcher canvas by selecting the **Hydrogen Generation** option from the **Options** top-level menu. Whilst this option is in effect hydrogen atoms are also automatically added to all new carbon substructures as they are added onto the *Sketcher* canvas. Note also that removal of hydrogens cannot be carried out, **Clear**, whilst the **Hydrogen Generation** option is on.
- Hydrogen atoms may also be drawn explicitly in the same way as any other type of atom.

Note: It is not possible to explicitly add hydrogens on substructures drawn with the molecule type *protein* (see Setting Molecule Types Section 1.3, page 120).

4.5 Setting Number of Connected Atoms

It is possible to specify the number of connections of an atom (i.e. the total number of atoms to which it is bonded).

To set the number of connections from a carbon atom:

- In **Draw** or **Select** mode, click on a carbon atom with the right-hand mouse button, pick **Number of connections from carbon** from the resulting pull-down menu, then select the number required from the second pull-down menu. **Unspecified** in this menu means that any number of connections is allowed.
- When setting the number of connections from carbon all atoms will be considered including hydrogens. It is only possible to specify this constraint on carbon atoms since the number of hydrogens cannot be safely inferred for other atom types.

To set the number of connections to non-hydrogen atoms:

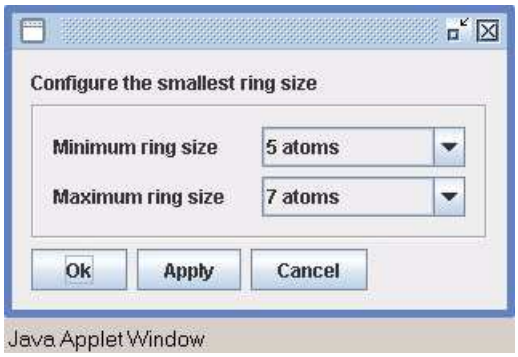
- In **Draw** or **Select** mode, click on an atom with the right-hand mouse button, pick **Number of connections to non-hydrogen atoms** from the resulting pull-down menu, then select the number required from the second pull-down menu. **Unspecified** in this menu means that any number of connections is allowed.
- This constraint can be set for any atom and will consider connected heavy atoms only (i.e. any connected hydrogen atoms will be ignored).

4.6 Defining Cyclic or Acyclic Atoms

- It is possible to specify that a particular atom must be cyclic (i.e. part of a ring) or, conversely, that it must be acyclic (i.e. not part of a ring).
- In **Draw** or **Select** mode, click on an atom with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, then select the required option from the second pull-down menu. **Unspecified** in this menu means the atom may be either cyclic or acyclic. If the atom is already part of a ring, the **Acyclic** option will not be active.
- For atoms which form part of a ring (i.e. cyclic atoms) it is also possible to specify the ring size. *Note:* When a particular atom is part of more than one ring the smallest ring will be considered when testing the constraint.
- To set a maximum limit on the ring size click on an atom with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Maximum smallest ring sizes**, then select the required option from the third pull-down menu.
- To set a minimum limit on the ring size click on an atom with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Minimum smallest ring sizes**, then

select the required option from the third pull-down menu.

- To specify an exact ring size click on an atom with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Exact smallest ring sizes**, then select the required option from the third pull-down menu.
- To specify a ring size range click on an atom with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Define Custom Ring Size**, then in the resulting pop-up window select the required minimum and maximum ring size, then hit **OK** (closes window) or **Apply** (leaves window open). The example below shows the setting required to specify that an atom must form part of a 5, 6, or 7-membered ring:



5 Bond Properties

5.1 Changing the Current Bond Type

- The current bond type determines the type of any new bond created when drawing. The current setting is shown at the bottom of the *Draw* window:

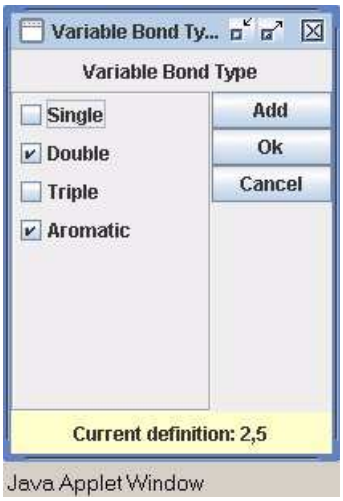
- The current bond type may be changed by clicking on this button and selecting from the resulting pull-down menu.
- Alternatively the bond type can be changed via the sketcher: right-click on the canvas (ensuring no bonds are selected), select **Set Bond Type** and pick the appropriate bond type from the resultant pulldown menu.
- **Any** means any covalent bond; bonds of this type are displayed as a dashed line.

5.2 Setting Variable Bond Types

- Bonds in a substructure may be variable, e.g. double or aromatic.
- The current bond type (see Changing the Current Bond Type Section 5.1, page 131) can be made

variable by clicking on the bond type button at the bottom of the *Draw* window. Select **Variable** from the pull-down menu, select the required bond types in the resulting pop-up window, then hit **OK** (closes window) or **Add** (leaves window open).

- The example below shows the setting required to create a variable bond type of double or aromatic:



5.3 Changing the Types of Existing Bonds

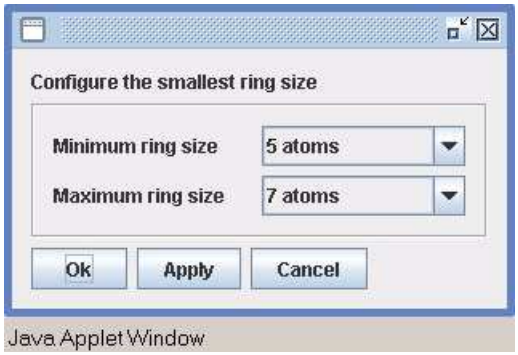
This can be done in several ways, including:

- In any mode, click on the bond with the right-hand mouse button and select **Set bond type...** from the resulting pull-down menu. Then select the required bond type.
- In **Draw** mode, change the current bond type (see Changing the Current Bond Type Section 5.1, page 131) and then click on the atom with the left-hand mouse button.
- In any mode, click on **Bond** in the top-level menu, select **Type** from the resulting pull-down menu and select the required bond type. The *Bond Property* pop-up appears, click on the bond or bonds to be changed with the left-hand mouse button and hit **Done**.

5.4 Defining Cyclic or Acyclic Bonds

- It is possible to specify that a particular bond must be cyclic (i.e. part of a ring) or, conversely, that it must be acyclic (i.e. not part of a ring).
- In **Draw** or **Select** mode, click on the centre of the bond with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, then select the required option from the second pull-down menu. **Unspecified** in this menu means the bond may be either cyclic or acyclic. If the bond is already part of a ring, the **Acyclic** option will not be active.
- For bonds which form part of a ring (i.e. cyclic bonds) it is also possible to specify the ring size. *Note:* When a particular bond is part of more than one ring the smallest ring will be considered when testing the constraint.

- To set a maximum limit on the ring size click on the centre of the bond with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Maximum smallest ring sizes**, then select the required option from the third pull-down menu.
- To set a minimum limit on the ring size click on the centre of the bond with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Minimum smallest ring sizes**, then select the required option from the third pull-down menu.
- To specify an exact ring size click on the centre of the bond with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Exact smallest ring sizes**, then select the required option from the third pull-down menu.
- To specify a ring size range click on the centre of the bond with the right-hand mouse button, pick **Cyclicity** from the resulting pull-down menu, followed by **Define Custom Ring Size**, then in the resulting pop-up window select the required minimum and maximum ring size, then hit **OK** (closes window) or **Apply** (leaves window open). The example below shows the setting required to specify that a bond must form part of a 5, 6, or 7-membered ring:



6 Using Substructure Templates

- Substructure drawing can be made easier by using templates, which are pre-drawn substructural fragments.
- To access the available templates, hit the **Other...** button in the *Template* sections on the left of the *Draw* window, then select either **Protein**, or **Ligand** templates in the resulting pull-down menu.
- Templates are available for ligands and for proteins:
  - For ligands, amino acid, steroid and saturated and unsaturated ring templates can be accessed.
  - For proteins, standard amino acid templates as well as modified amino acid templates (e.g. phosphorylated tyrosine) are provided.
- Once selected, the template should be moved into the desired position using the mouse. To scale the template move the mouse while keeping the *Shift* button depressed, to rotate move the mouse while keeping the *Control* button depressed. To fuse the template with an existing substructure (of the same molecule type) move it towards the substructure. Fusion is achieved by overlapping one or more atoms in the template with one or more atoms in the existing substructure (indicated

by the overlapped atoms being highlighted). Once a template is in the desired position click the left mouse button to load it into the sketcher.

- Alternatively, to load a template, hit **File** in the top-level menu of the *Draw* window, followed by **Import Template** and select a template from the resulting menus.
- Note that protein and ligand templates contain implicit H atoms (pass the mouse cursor over atoms to see how many H atoms are bonded to them). Implicit H atoms are present to resolve any ambiguities that may arise (e.g. glycine has its alpha carbon protonated, otherwise glycine would match all other amino acids). H atoms can be removed from **All Atoms** or **Selected Atoms** of the template if required, via **Atom**, **Hydrogens**, **Clear** (see Addition of Hydrogen Atoms Section 4.4, page 129).

7 Substructure Display Conventions

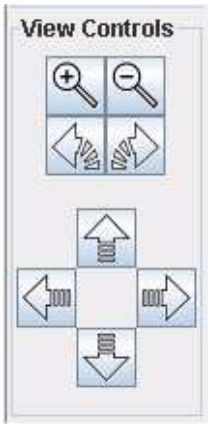
Most of the conventions and symbols used in displaying substructures are obvious. Those that are not include:

- An atom whose symbol begins with the letter *V (V1, V2*, etc.) has a variable element type. Positioning the cursor over the atom will display a help message giving further details.
- A superscript beginning with the letter *T* indicates the total number of connected atoms (including hydrogen atoms), e.g. *T4* indicates that the atom must be 4-coordinate.
- A superscript beginning with the letter *X* indicates the total number of connected heavy atoms, e.g. *X4* indicates that the atom must be connected to 4 non-hydrogen atoms.
- The letter *a* indicates acyclic; *c* indicates cyclic. Any further ring size constraints are indicated in square brackets, e.g. *c[<7]* indicates the atom must form part of a ring with a maximum ring size of no more than 6 members.
- If an atom is surrounded by a circle, it is close to, or on top of, another atom. Change to **Select** mode, select the atom, and move it away.
- Atom labels for carbon atoms may be hidden by selecting the top-level menu **Options** and checking the **Hide Carbons** box.
- The font size used for labels can be changed by selecting **Font** from the **View** menu and picking the appropriate **Increase/Decrease/Default** font size.

8 Moving, Scaling, Rotating and Duplicating Substructures

8.1 Moving Atoms

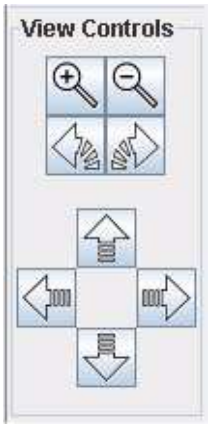
- Select (see Selecting Atoms Section 2.7, page 123) the atom(s) to be moved.
- Press the left-hand mouse button, and move the cursor while keeping the button depressed. Release the left-hand mouse button at the desired position.
- To move the complete query (i.e. the entire contents of the sketcher window) use the scroll buttons in the *View Controls* section to the left of the *Draw* window:



- To re-centre the complete query in the sketcher window select **View** from the top-level menu, followed by **Recentre View** from the resulting pull-down menu.
- To automatically move (and rescale) the complete query such that it will fit into the sketcher window select **View** from the top-level menu, followed by Auto**Fit** from the resulting pull-down menu.

8.2 Scaling Queries

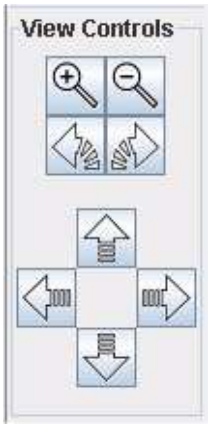
- It is only possible to scale complete fragments, not a collection of atoms which form part of a fragment. Select the fragment(s) to be scaled and then clicking on one of the corners drag the mouse till the fragment is the appropriate size
- In any mode, use the mouse scroll wheel to adjust the scale of the contents of the drawing area. Alternatively, use the zoom buttons in the *View Controls* section to the left of the *Draw* window:



- To automatically resize the complete query to the maximum, minimum, or default zoom level select **View** from the top-level menu, **Zoom** from the resulting pull-down menu, then select the required option from the subsequent menu.
- To automatically rescale (and move) the complete query such that it will fit into the sketcher window select **View** from the top-level menu, followed by Auto**Fit** from the resulting pull-down menu.

8.3 Rotating Queries

- It is only possible to rotate complete fragments, not a collection of atoms which form part of a fragment.
- Select (see Selecting Atoms Section 2.7, page 123) the fragment(s) to be moved.
- While keeping the *Control* button depressed use the left-hand mouse button to rotate the fragment(s). Release the left-hand mouse button at the desired position.
- To move the complete query (i.e. the entire contents of the sketcher window) use the rotate buttons in the *View Controls* section to the left of the *Draw* window:

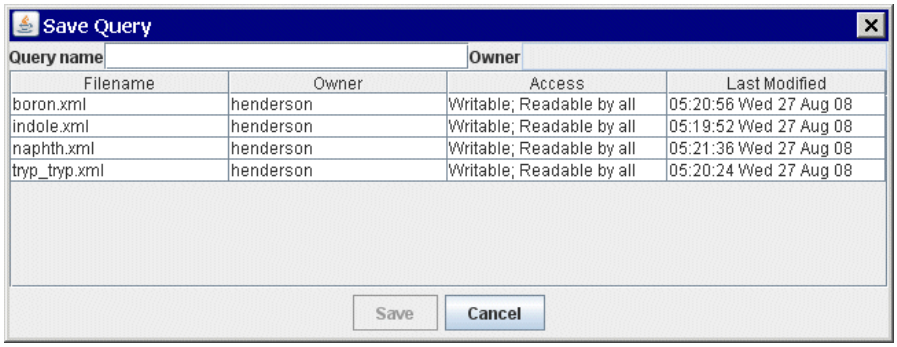


8.4 Duplicating Substructures (Copy and Paste)

- To make a copy of all or part of a substructure, select (see Selecting Atoms Section 2.7, page 123) the atoms and bonds to be copied.
- Click on a blank point in the white area with the right-hand mouse button and select **Copy,** or hit **Edit** in the top-level menu and **Copy** in the resulting pull-down menu. Alternatively use the keyboard shortcut CTRL-C.
- A copy of the selected substructure (or part substructure) will appear in the sketcher. The substructure copy should be moved into the desired position using the mouse. To scale the substructure move the mouse while keeping the *Shift* button depressed, to rotate move the mouse while keeping the *Control* button depressed. To fuse the substructure copy with an existing substructure (of the same molecule type) move it towards the substructure. Fusion is achieved by overlapping one or more atoms in the substructure with one or more atoms in the existing query (indicated by the overlapped atoms being highlighted). Once a copied substructure is in the desired position click the left mouse button to load it into the sketcher.

9 Reading, Saving and Deleting Queries

- To save a query set up in the sketcher, select **File** in the top-level menu, then **Save Query** in the resulting pull-down menu. This will open the *Save Relibase+ query* pop-up window:



- Enter a query name and hit the **Save** button. To make the query readable by all users select **Yes** in the subsequent window.
- Saved queries can be read back into the drawing area by selecting **File** in the top-level menu, then **Read Query** in the resulting pull-down menu and selecting the saved query from those appearing in the resultant dialog box.
- To delete a query select **File** in the top-level menu, then **Load/Delete Query** in the resulting pull-down menu, select the query from those appearing in the resultant dialog box and hit **Delete**. It is not possible to delete queries that are owned by other users.
- Queries may be pasted from third-party sketchers in MOL/SDFile format via the **Paste Query from System Clipboard** in the **Edit** menu.

Note: for ISIS/Draw it is necessary to first enable the **Copy Mol/Rxnfile to the Clipboard** option under **ISIS/Draw**, **Settings**, **General**.

10 Geometric Objects

Geometric objects may be defined when drawing a substructure in the *Draw* window. These objects can then be used for computing geometric parameters, e.g. the distance between two centroids.

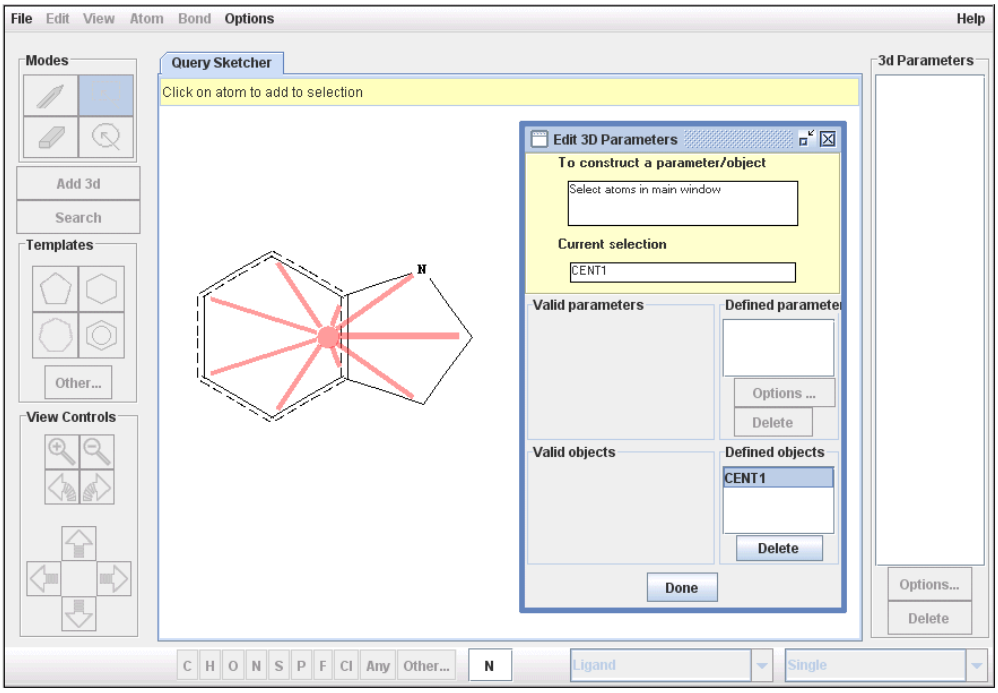
10.1 Valid Geometric Objects

Valid objects are:

- Centroids
- Vectors
- Planes

10.2 Defining Geometric Objects

- Open up the *Geometric Parameters* dialogue box by clicking on the **Add 3d** button in the *Draw* window. This button will only be active when there is a substructure in the drawing area.
- Select the atoms or existing objects in the *Valid objects* list that are needed to calculate the new object by clicking on them with the left-hand mouse button (click again on an atom to deselect).
- As the number of selected atoms varies, the dialogue box will list the *Valid objects* that can meaningfully be defined.
- Hit the appropriate **Define** button in the dialogue box, e.g. next to the word *Centroid* to define a centroid.
- The defined object is listed in the box labelled *Defined Objects*.
- In the example below, the centroid of an indole substructure has been defined:



10.3 Displaying Geometric Objects in the Draw Window

- A defined object may be displayed by clicking on its name in the *Defined Objects* list. This may be found in the *Geometric Parameters* dialogue box (opened by hitting the **Add 3d** button).

10.4 Deleting a Geometric Object

- An object may be deleted by opening the *Geometric Parameters* dialogue box (click on the **Add 3d** button), clicking on the object name in the *Defined Objects* list, and then hitting the **Delete** button underneath this list.

11 Geometric Parameters

- Geometric parameters can be defined when drawing a substructure in the drawing window.
- Histograms of these parameters can then be viewed after the search has been run.
- Geometric parameters can also be used to set up 3D substructure searches (e.g. a search for a substructure in which a distance has been constrained to a particular range).

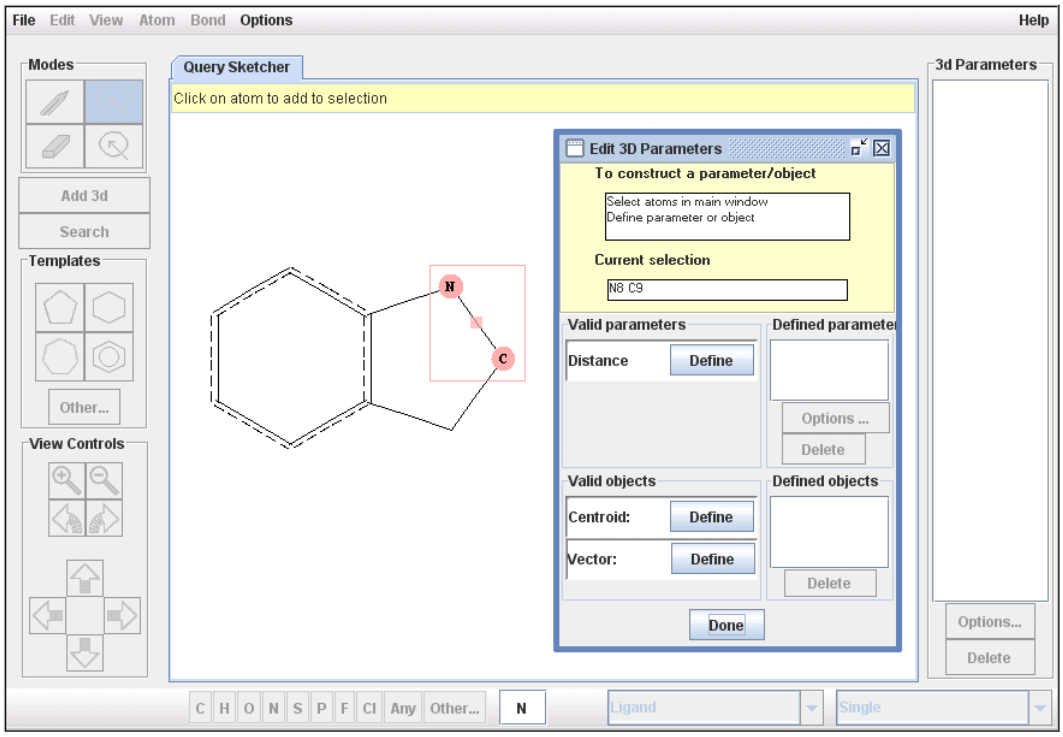
11.1 Valid Geometric Parameters

Valid geometric parameters are:

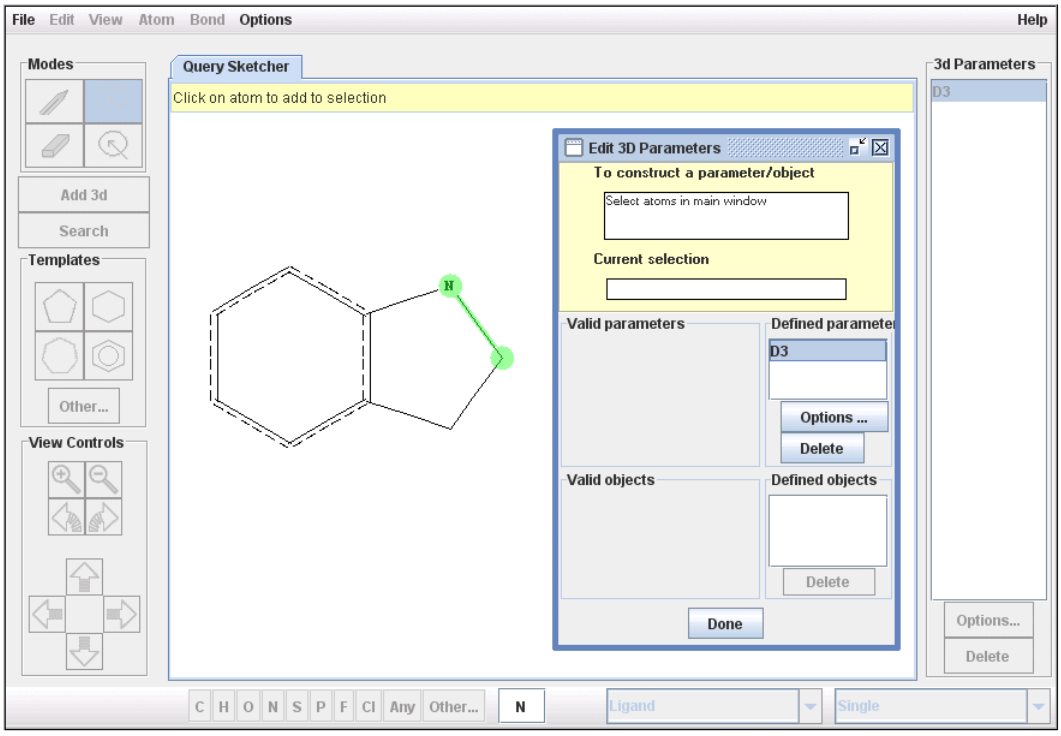
- Distances between atoms and/or objects; the atoms do not need to be bonded to each other.
- Angles between atoms and/or objects; the atoms do not need to be bonded to one another.
- Torsion angles involving atoms and/or objects; the atoms do not need to be bonded to one another.

11.2 Defining Geometric Parameters Involving Atoms

- Geometric parameters must be explicitly defined in the *Draw* window in order to be displayed as a histograms after the search has been run.
- Open up the *Geometric Parameters* dialogue box by clicking on the **Add 3d** button in the *Draw* window. The dialogue box can only be opened when there is a substructure in the white drawing area.
- Select the atoms that are needed to calculate the required parameter by clicking on them with the left-hand mouse button (click again on an atom to deselect).
- As the number of selected atoms varies, the dialogue box will list the parameters that can meaningfully be defined. In the example below, two atoms have been selected so it is possible to define the distance between them:



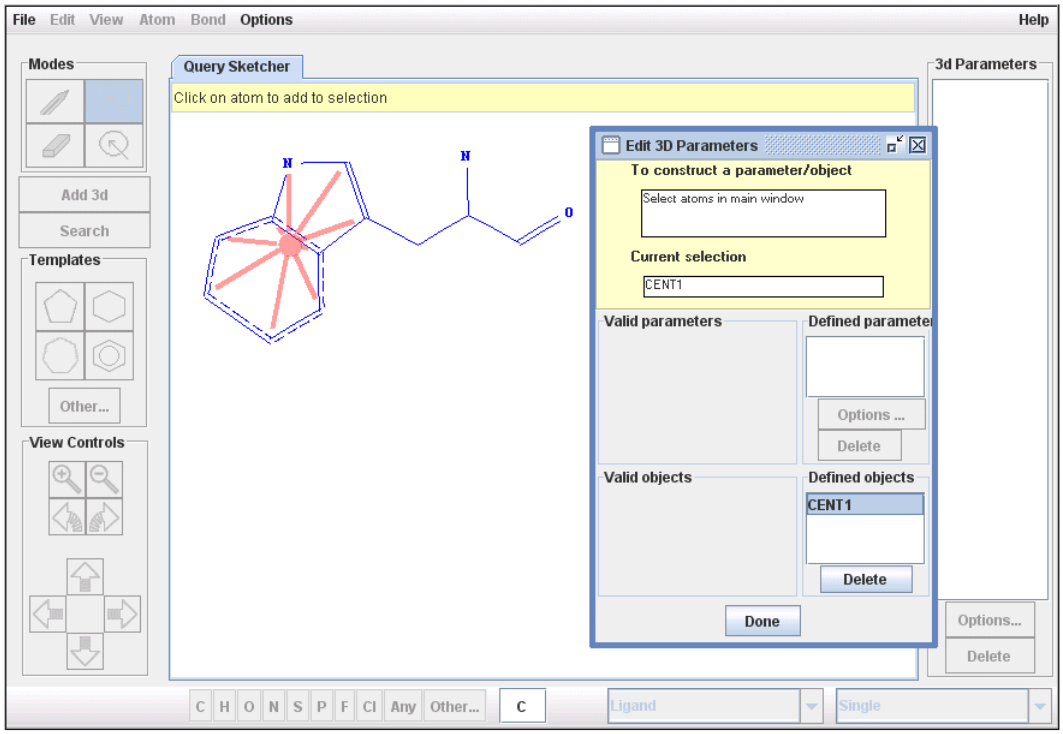
- Hit the appropriate **Define** button in the list of *Valid Parameters*, e.g. hit the **Define** button next to the word *Distance* to define an interatomic distance (atoms do not need to be bonded to one another).
- The defined parameter will be listed in the box labelled *3d Parameters* in the top-right hand corner of the *Draw* window. In the example below, the N...C distance has been defined and named *D1* by default:



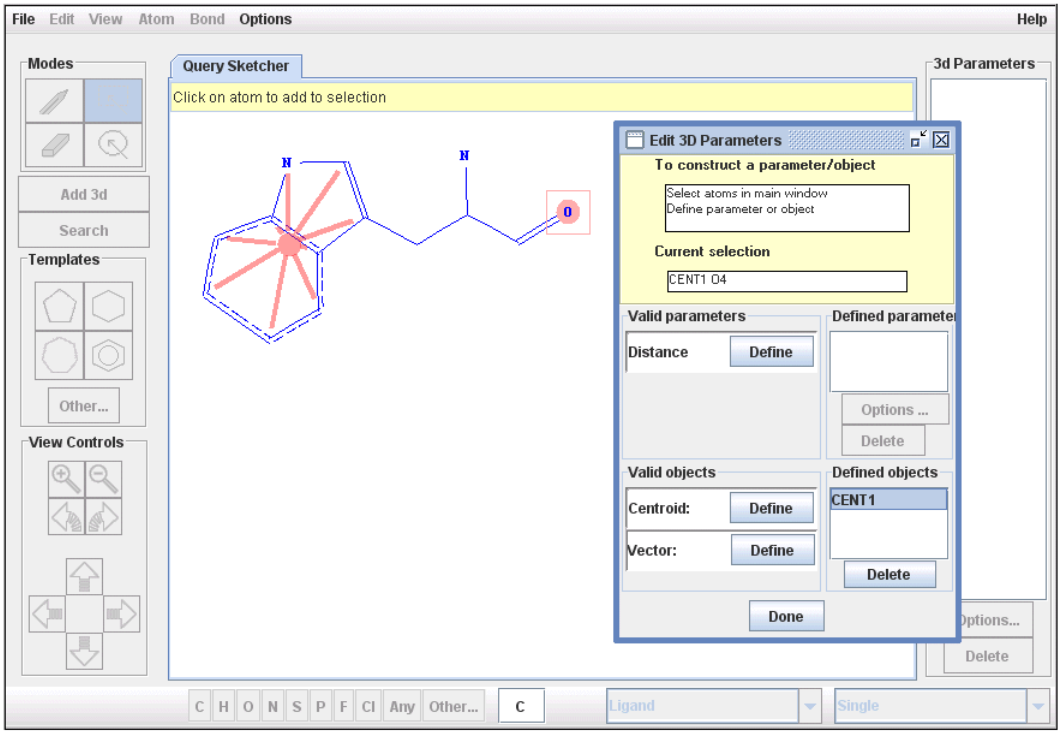
- Once a parameter has been defined, its value can be constrained (see Applying Constraints Section 12, page 144).
- You can continue to define other parameters. Once all parameters of interest have been defined hit **Done** to close the *Geometric Parameters* dialogue box.

11.3 Defining Geometric Parameters Involving Objects

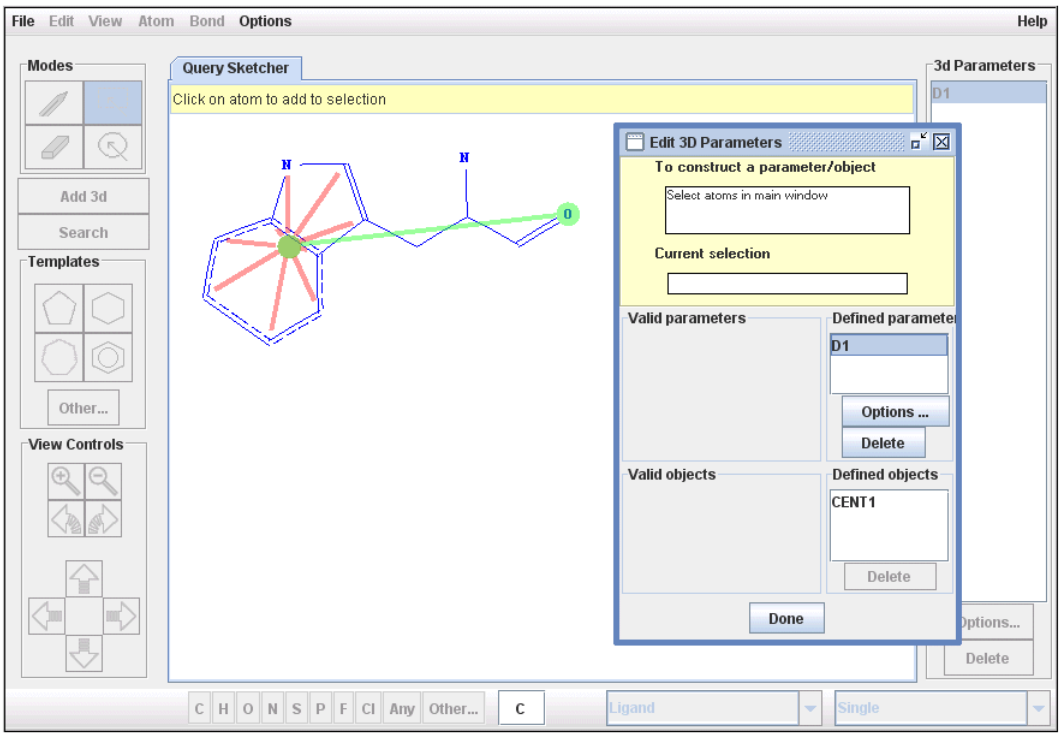
- The procedure is exactly the same as for parameters involving only atoms (see Defining Geometric Parameters Involving Atoms Section 11.2, page 140), except that objects are picked from the *Defined Objects* list.
- For example, to specify the distance between a centroid and an atom, first create the centroid (see Defining Geometric Objects Section 10.2, page 138):



- Then, select the atom by clicking on it, and the centroid by clicking on its object name in the list of *Defined Objects* (*CENT1* in the above example):



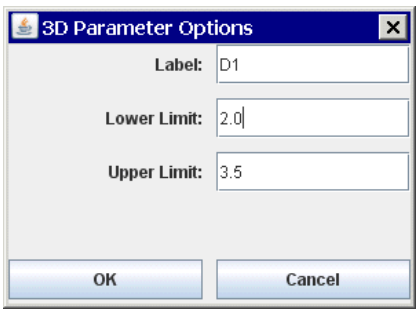
- Then hit the relevant **Define** button (next to the word *Distance* in the above example).



- Once a parameter has been defined, its value can be constrained (see Applying Constraints Section 12, page 144).
- You can continue to define other parameters. Once all parameters of interest have been defined hit **Done** to close the *Geometric Parameters* dialogue box.

11.4 Renaming Geometric Parameters

- By default, distances are named *D1, D2,* etc.; angles *A1, A2,* etc.; torsions *T1, T2,* etc.
- To rename a parameter select it in the *3D Parameters* list (top right-hand corner of *Draw* window), then hit the **Options...** button underneath this list.
- In the resulting pop-up, type the new name into the *Label* input box, e.g.



- Alternatively, you can rename a parameter immediately after creating it by hitting the **Options...** button in the *Geometric Parameters* dialogue box (opened by hitting **Add 3d**).

11.5 Displaying Geometric Parameter in the Draw Window

- A defined parameter may be displayed by clicking on its name in the *3d Parameters* list (top-right hand corner of *Draw* window).

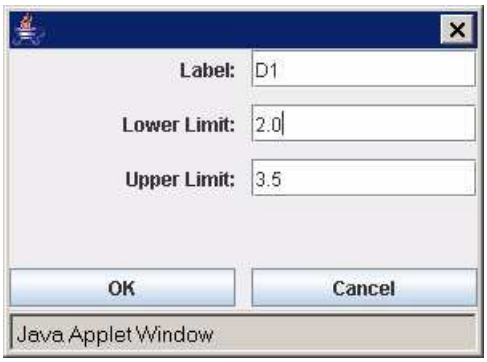
11.6 Deleting a Geometric Parameter

- A parameter may be deleted by clicking on its name in the *3d Parameters* list (top-right hand corner of *Draw* window) and then clicking on the **Delete** button underneath this list.

12 Applying Constraints

12.1 Geometric Constraints

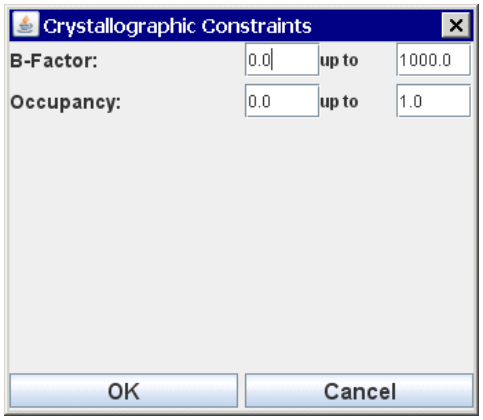
- 3D substructure searches are performed by defining relevant geometric parameters (see Defining Geometric Parameters Involving Atoms Section 11.2, page 140) and constraining their values. To constrain the value of a defined parameter:
- If you have just defined the parameter, so that the *Geometric Parameters* dialogue box is already open, hit the **Options...** button.
- If the *Geometric Parameters* dialogue box is not already open, select the parameter you wish to constrain in the *3d Parameters* list (top right-hand corner of the *Draw* window) and hit the **Options...** button underneath this list.
- In the resulting dialogue box, enter the required *Lower Limit* and *Upper Limit* values. In the example below, a distance *D1* has been constrained to values between *2.0* and *3.5*Å:



- In the case of torsions, the chosen limits can be constrained to be within the range *0* to *+360* degrees, or *-180* to *+180* degrees. Use the **Change Torsion Range** check box to change the convention used. The same range must be used for all torsion angles in a search (The **Change Torsion Range** box is inactive if more than one torsion angle has been defined).

12.2 Crystallographic Constraints

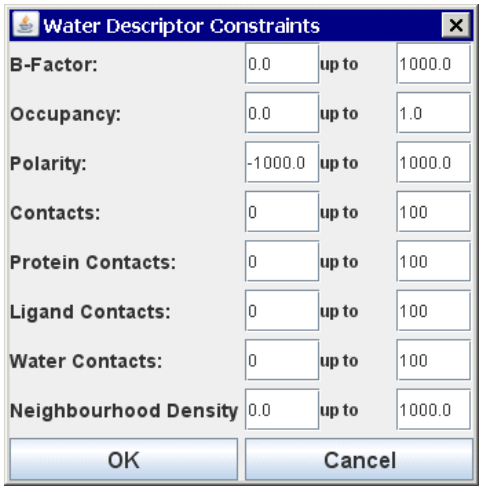
- Crystallographic constraints can be used to constrain the properties of atom(s) when setting up 2D or 3D queries in the *Draw* window.
- To constrain an atom, right click on the atom and select **Crystallographic Constraints...** from the resulting menu, this will launch the *Crystallographic Constraints* dialog box:



- The range for the allowed values of crystallographic B-factor and occupancy can be set.

12.3 Water Descriptor Constraints

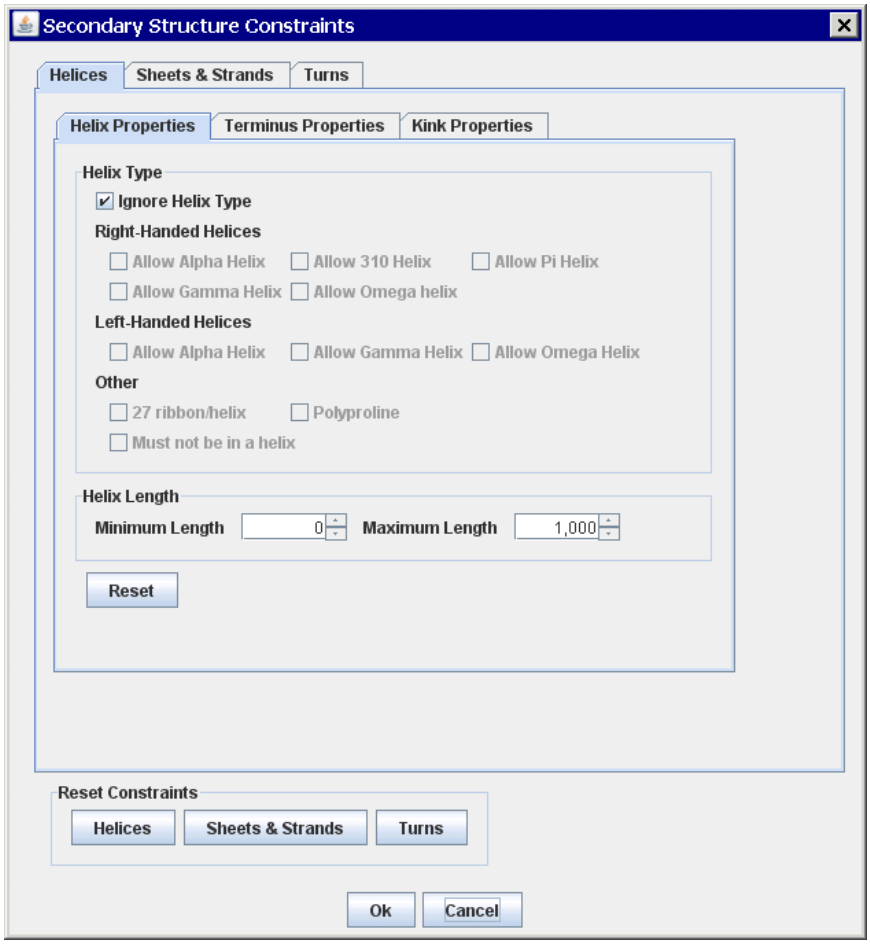
- Water descriptors (see Water Molecule Descriptors Section 5.6, page 4) can be used to constrain the properties of water molecules when setting up 3D queries in the *Draw* window.
- To constrain a water molecule ensure its *Molecule Type* is set to *Water* (see Setting Molecule Types Section 1.3, page 120), then right click on the oxygen atom of the water and pick *Constrain Atom* from the pulldown menu. This will launch the *Water Descriptor Constraints* dialog box:



- The range for the values of crystallographic B-factors, polarity, number of contacts, and neighbourhood density can be set.

12.4 Secondary Structure Constraints

- The secondary structure in the protein (see Secondary Structure Information Section 5.8, page 10) can be used to constrain protein substructure searches.
- To constrain a protein atom(s) or residue, ensure its *Molecule Type* is set to *Protein* (see Setting Molecule Types Section 1.3, page 120), right click on the selected atom or area, then select **Secondary Structure Constraints** from the pulldown menu. This will launch the *Secondary Structure Constraints* dialog box:



13 Defining Secondary Structure Elements

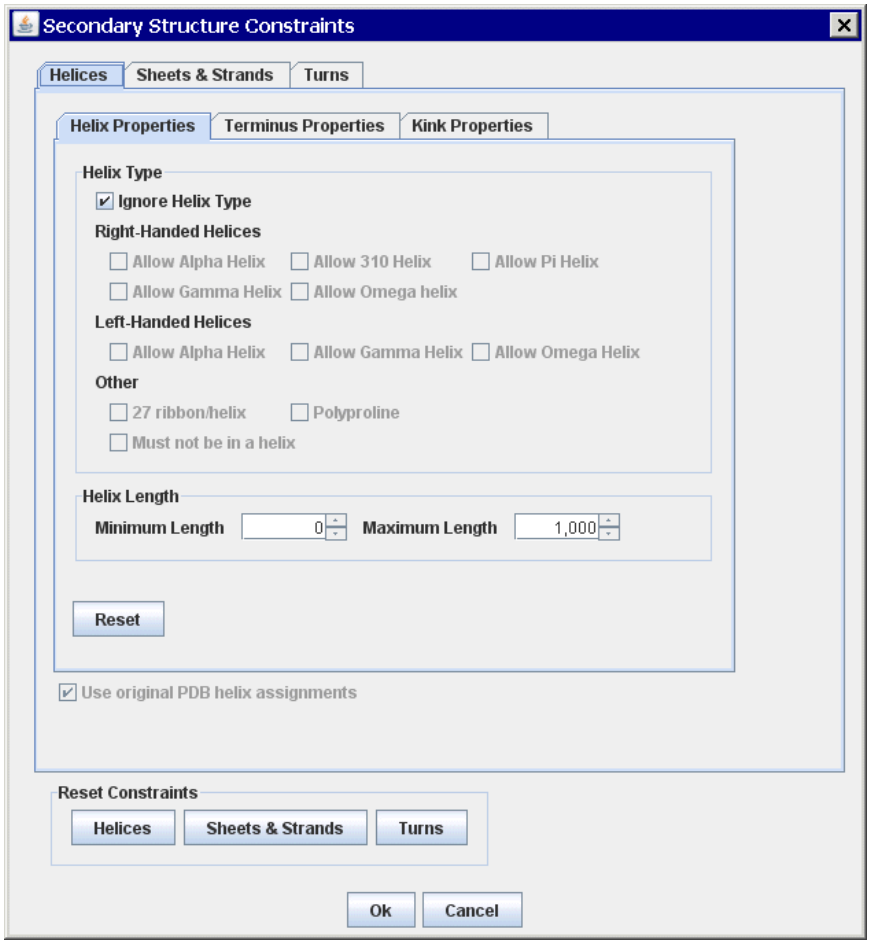
13.1 Overview

- Secondary structure elements can be defined in 2D and 3D sketcher searches. This can be combined with other search tools to provide powerful restricted searches of protein amino acids in given loop conformations.

- Secondary structure assignments can be accessed via protein and ligand information pages and viewed in AstexViewer (see Viewing Secondary Structure Assignments Section 4.6.6, page 38).
- An overview of the methodology involved in compiling the secondary structure module is provided elsewhere (see Secondary Structure Information Section 4.6, page 35).

13.2 Constraining a Protein Residue to be in a Particular Secondary Structure Element

- After sketching an amino acid in the sketcher (see Using Substructure Templates Section 6, page 133), right click on any atom in the amino acid and select **Secondary Structure Constraints** from the resultant pull-down menu. This will launch a *Secondary Structure Constraints* dialogue window.



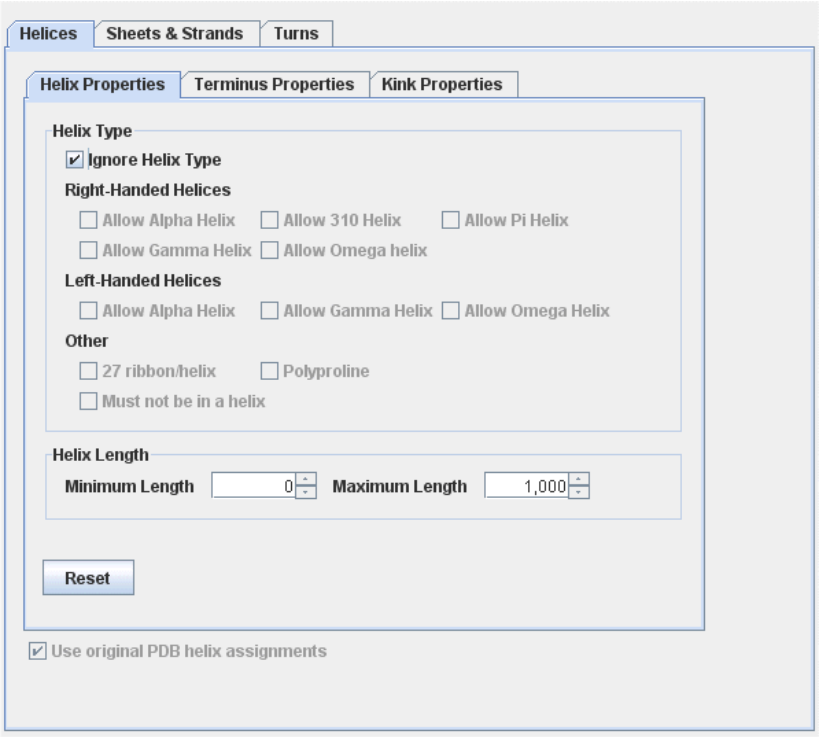
- Using this dialog it is possible to constrain the secondary structure of the amino acid to be in either a sheet, a helix or a turn.
- The dialog is separated into 3 major sections:
  - Helices (see Defining Helix Properties Section 13.3, page 148).
  - Sheets & Strands (see Defining Sheet and Strand Properties Section 13.4, page 150).
  - Turns (see Defining Turn Properties Section 13.5, page 152).

- Use the **Reset** button in each of the separate sub-section above to clear the pane that is currently on view and return it to its original settings.
- Use the **Reset Constraints** buttons at the bottom of the dialog to reset all constraints of a given class to the default.

13.3Defining Helix Properties

- Within the *Helix Properties* pane it is possible to constrain on the basis of the properties of a particular helix (type and length).
- The Helices tab is subdivided into three panes:
  - Helix Properties.
  - Terminus Properties.
  - Kink Properties.
- By default, original PDB assignments of helices are when one specifies a helix constraint. This can be changed to use the *SHAFT* assignment by unchecking the **Use original PDB helix assignments** check box.

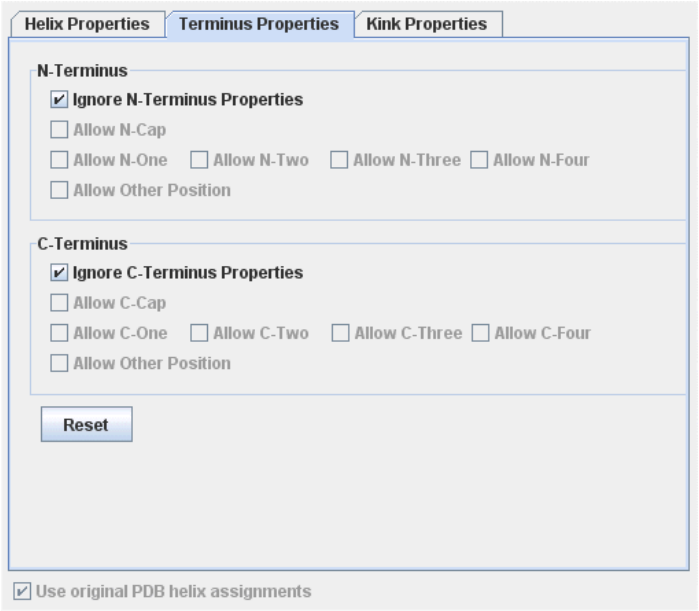
Helix Properties



- To define helix properties, deactivate the **Ignore Helix Type** check box. It will then become possible to select *Right-Handed Helices*, *Left-Handed Helices* and *Other* properties via their individual check boxes.

- The minimum and maximum helix length can be defined in the *Helix Length* section of the panel.
- Use the **Reset** button to return the *Helix Properties* settings to their original display.
- Note: the SHAFT assignment only contains right-handed helices (3_{10}, a and p helices).

Terminus Properties



- Use this tab to specify the properties of a residue with respect to the N or C terminus.
- The *N-cap* and *C-cap* residues are the residues at either end of a helix, with the N-one, N-two etc. being steps along from the N-cap residue (so the N-one would be the residue in the helix adjacent to the N-cap residue). Similarly, the C-one, C-two residues follow a similar pattern.
- Dectivate the **Ignore N-Terminus Properties** or **Ignore C-Terminus Properties** check box to enable and define N-Terminus and C-Terminus properties.
- Use the **Reset** button to return the settings to their original display.
- Note: the capping residues for helices are: N_{cap} to N_{two} and C_{two} to C_{cap} for 3_{10} helices, N_{cap} to N_{three} and C_{three} to C_{cap} for a-helices, N_{cap} to N_{three} and C_{three} to C_{cap} for p-helices.

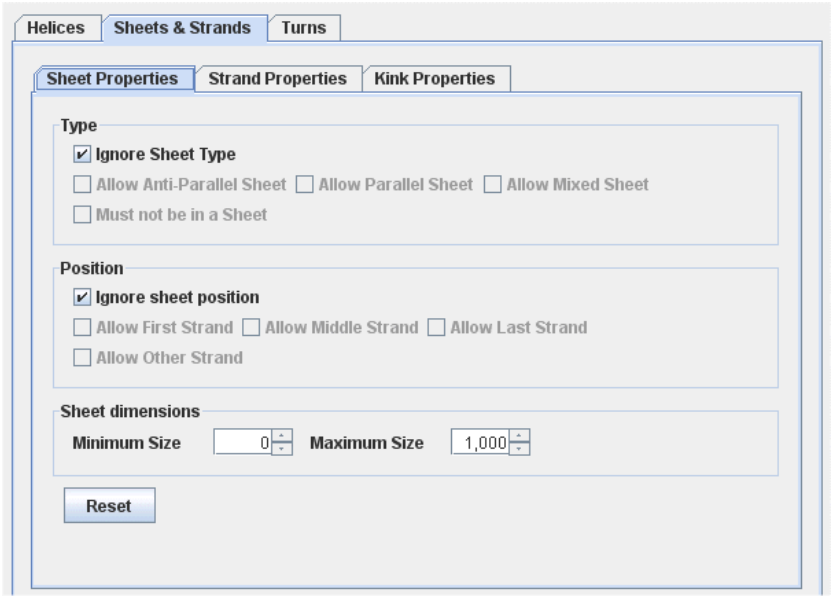
Kink Properties

- Use the *Kink Properties* tab to define and search for amino acids in kinks.
- Kinks are points in helices where the direction helical vector (mapping along the centre of the helix from the N-cap to the C-cap) change. The secondary structure module allows such kinks to be searched for, either where one change of direction occurs, or where two adjacent changes of direction occur.
- Deactivate the **Ignore kink type** check box to enable and select kink properties to be searched for.
- Use the **Reset** button to return the settings to their original display.
- Note: the kink properties are assigned to the midpoint of the Ca-atoms of four adjacent residues. You can define one of these Ca-atoms in the case of one kink and one Ca-atoms that is involved in two adjacent kinks in the other case.

13.4Defining Sheet and Strand Properties

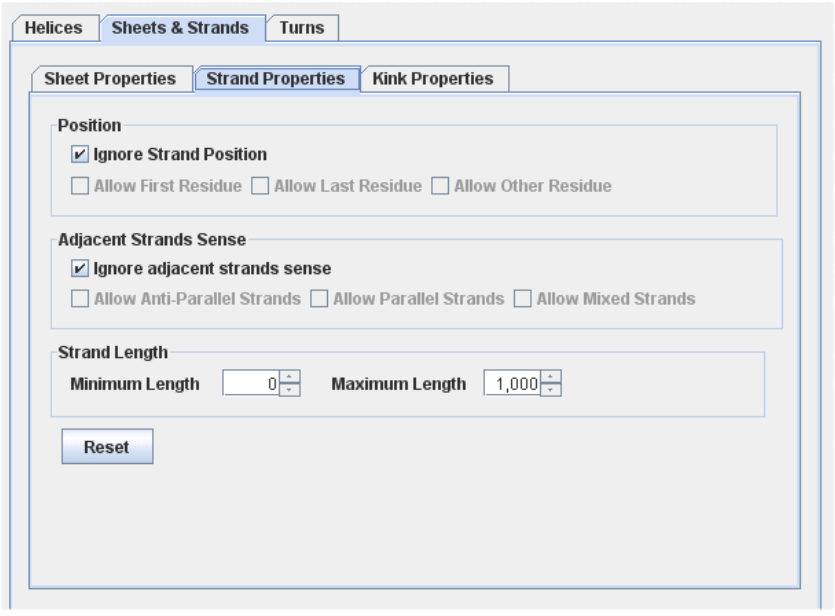
- Within the *Sheet & Strand Properties* pane it is possible to constrain residues to lie in sheets, strands or kinks with given properties.
- The *Sheets & Strands* tab is subdivided into three panes:
  - Sheet Properties.
  - Strand Properties.
  - Kink Properties.

Sheet Properties



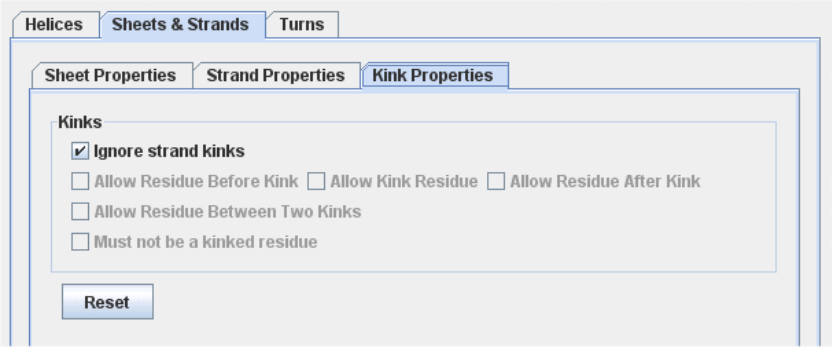
- Sheets can be constrained so that one searches for types of sheet, for specific strands in a given sheet. The *first strand* in a sheet is defined as the first strand that one comes to along the sequence as one traverses from the N-terminus residue of the sequence to the C-terminus.
- Specific sheet *Types* and *Positions* can be specified by deactivating the **Ignore Sheet Type** and **Ignore sheet position** check boxes. Sheet types can be constrained so that a search only returns sheets that are *parallel*, *anti-parallel* or *mixed*.
- Sheet sizes can be constrained so that sheets containing a defined number of residues are searched for.

Strand Properties



- From within the *Strand Properties* tab it is possible to constrain a search to look at individual strands that may or may not be in a sheet.
- Properties concerning the position of the strand and the sense of the adjacent strand can be defined by deactivating the **Ignore Strand Position** and **Ignore adjacent strands sense** check boxes and selecting from the resultant options.
- The strand length can be constrained to contain a user-defined number of residues.
- Use the **Reset** button to return the settings to their original display.

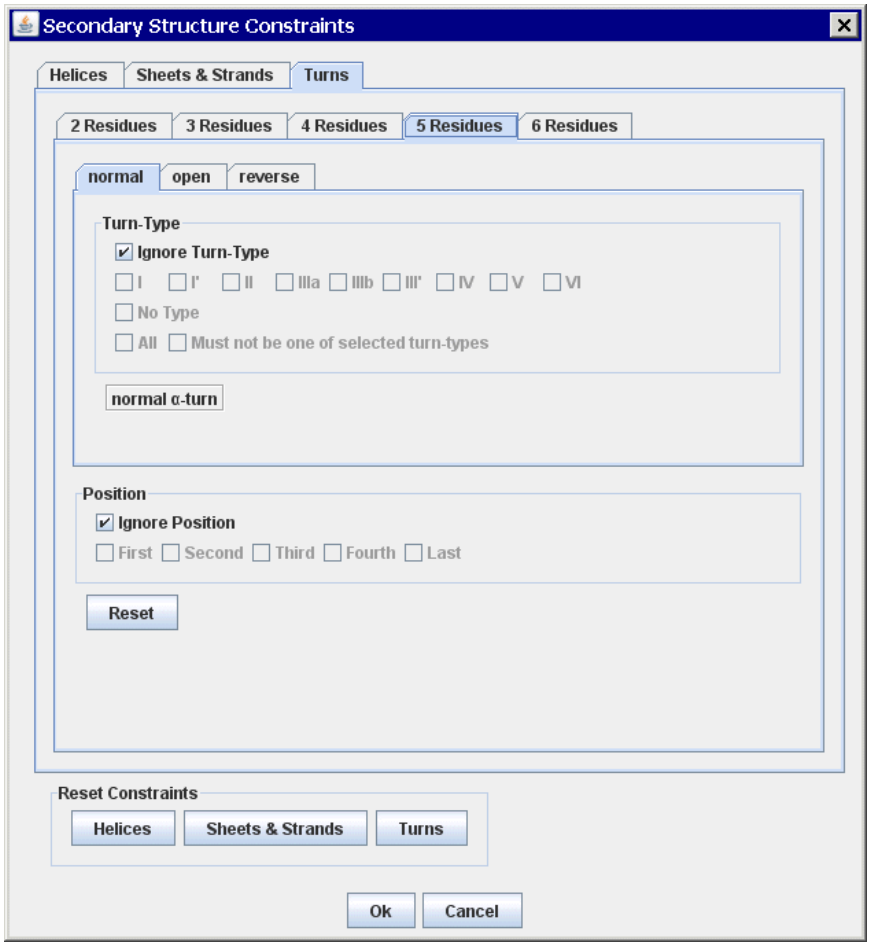
Kink Properties



- Kinks in a sheet can be defined from within the *Kink Properties* tab. A kink in a sheet is similar to a kink in a helix: each sheet has a directional vector associated with each strand. If this vector changes within a given strand the residue where the directional change occurs is said to be kinked. The secondary structure module allows for searching for residues that are in kinks or adjacent to kinks.
- To define kink properties, deactivate the **Ignore strand kinks** tick box and select from the resultant options.
- Use the **Reset** button to return the settings back to their original display.

13.5Defining Turn Properties

- The secondary structure module also provides the ability to search for specific turns in PDB entries. For turn searches, automatic assignments from the following publication are always used:
  Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions.
  O. Koch, G. Klebe, *Proteins: Structure, Function, and Bioinformatics*, 74, 353-367, 2008. [DOI: 10.1002/prot.22185]
  All types of turn described in the above publication can be searched.



- Turns are irregular secondary structure elements with a hydrogen bond or a specific Ca-Ca distance between the first and the last residue. Turns can be up to six residues in length thus a tab exists for each turn length.
- Within each turn length tab, further tabs are available based on the types of turns available for the turn length: *2-Residue* turns can only be the *reverse* type; *3-Residue* can be *normal* or *reverse*; *4-Residue*, *5-Residue* and *6-Residue* turns can be *normal*, *open* or *reverse*.
- To select turn types, deactivate the **Ignore Turn-Type** check box to activate and select the allowed options.
- Each pane allows the user to specify the relative position that a hit residue occupies in the turn. Simply deactivate the **Ignore Position** check box and select from the resultant options.
- Use the **Reset** button to return the window to its initial display.

CHAPTER 6: CREATING IN-HOUSE DATABASES

1 **Introduction** (see page 157)
2 **Overall workflow** (see page 157)
3 **Ligand templates** (see page 159)
4 **Synonyms file** (see page 160)
5 **Customising the processing requirements** (see page 161)
6 **Structure factors and electron densities** (see page 163)
7 **Processing structures using the web-based GUI** (see page 164)
8 **Processing structures using the command line** (see page 168)

1 Introduction

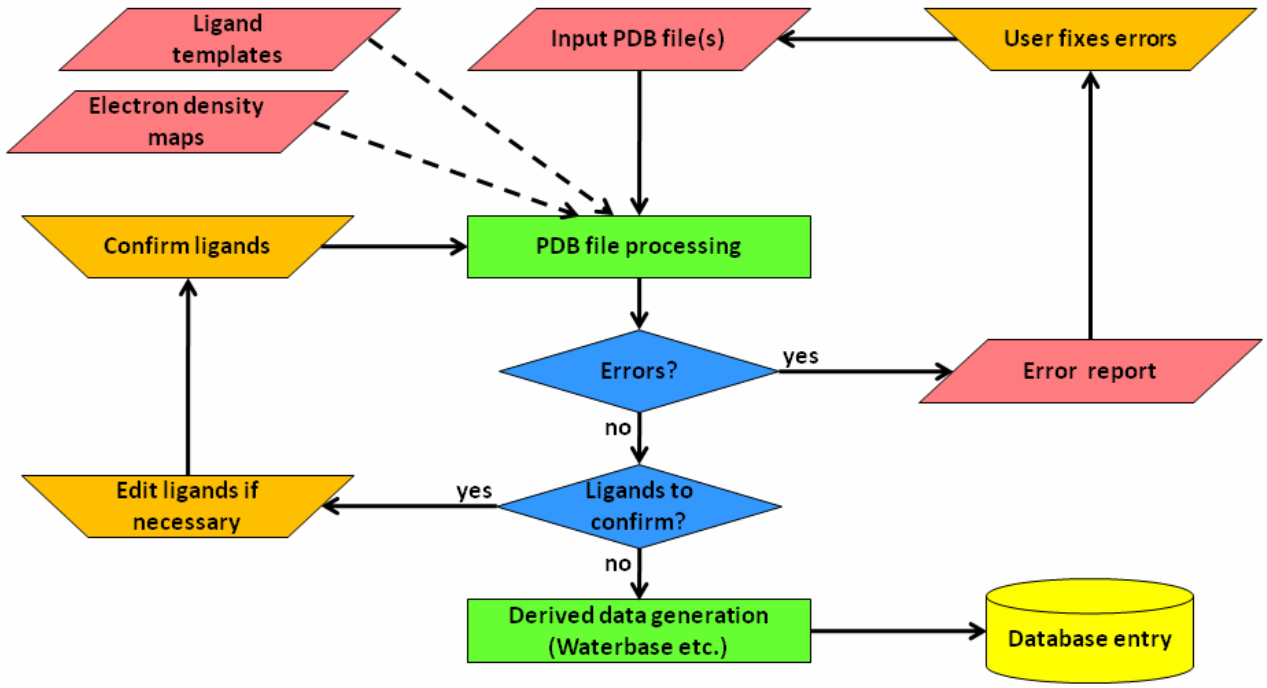
- Relibase+ provides a complete solution for storing and managing both public and proprietary structural data. This document describes the process of uploading in-house structures into Relibase+ by translating PDB files into Relibase+ database entries.
- The Relibase+ data processing system has been designed to be highly flexible and easy to use. The main reason for this is that there are many different scenarios for uploading protein structures into Relibase+. A protein crystallographer uploading a single structure has got different requirements to a molecular modeller uploading a backlog of 3000 structures. Both these scenarios are catered for in the Relibase+ data processing system.

2 Overall workflow

- Conceptually data processing can be broken down into two steps:

1. Ensure the PDB file is in an acceptable format, required as in-house structures may deviate from the PDB file format.
2. Ensure the ligand atom and bond typing are acceptable, required because the PDB file format does not contain information on ligand bond types.

- Several features have been implemented to make the above tasks as painless as possible.
- Data processing converts PDB files into Relibase+ database entries. The first step of the data processing parses the PDB file and detects any errors or missing fields. These are reported back to the user who is prompted to fix them. After any errors have been corrected the user will have to confirm any ligands that do not already have templates. Ligand templates are required to ensure correct atom and bond typing. Once all ligands in the protein structure have templates the data processing continues, calculating additional information such as crystal packing around the binding sites and cavities in the protein. Finally, the data are added as a new entry to the Relibase+ database.



- Several features have been implemented to minimise the amount of manual intervention required:

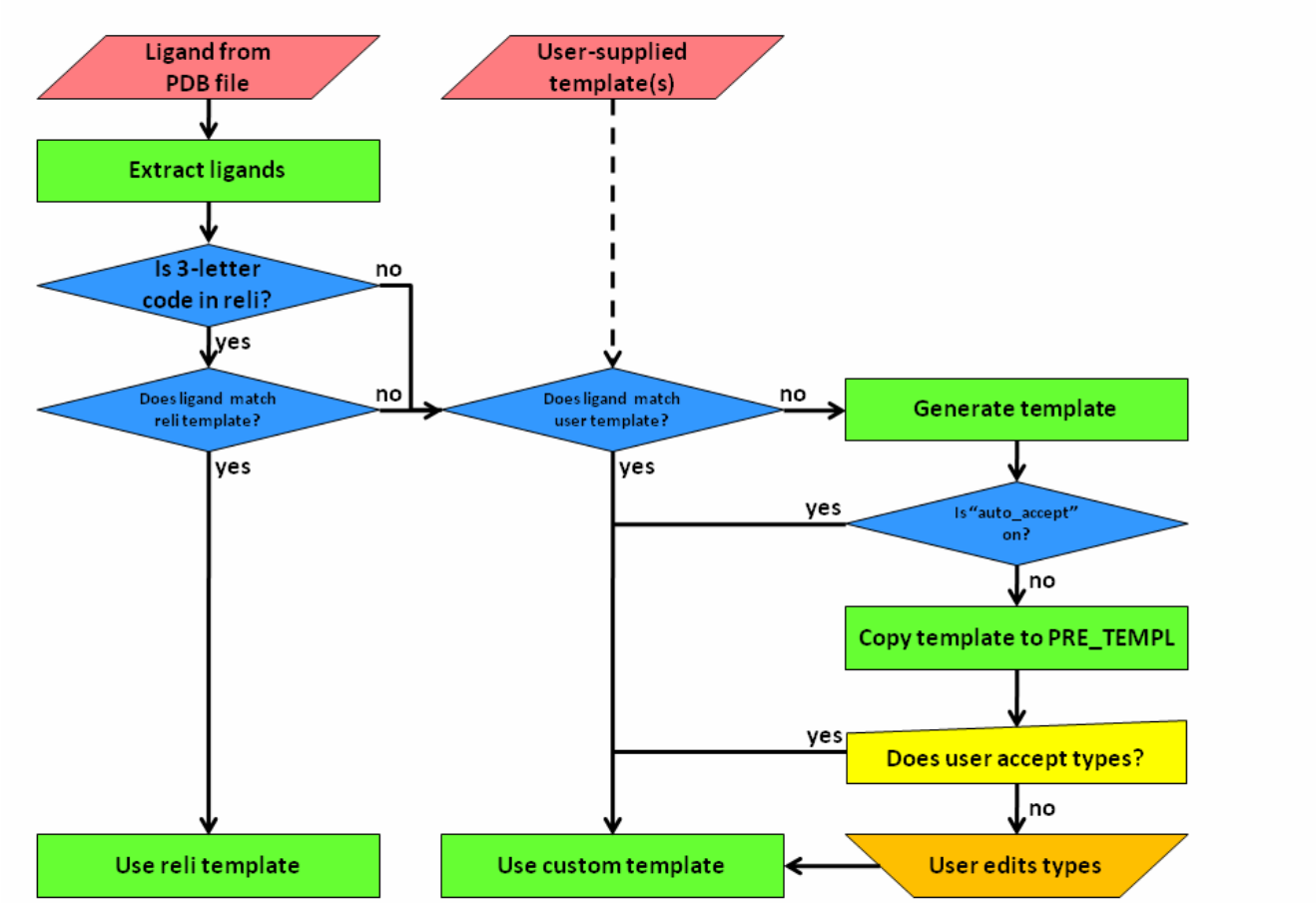
1. The strictness of data processing, in terms of which fields are required and the syntax allowed in the PDB file, can be customised in the `relibase_processing.conf` file. So for example if none of your in-house structures contain an `AUTHOR` record you can set the `author_required` flag to `false`. Alternatively, if you insist on the `AUTHOR` record being present in your in-house structures you can set the `author_required` flag to `true`.
2. If you do want to avoid having to validate ligand templates mid-processing there is an option to supply ligand templates with the PDB file at the beginning of the process. This option is particularly powerful if you want to process a backlog of thousands of in-house structures.
3. Another powerful feature when dealing with legacy structures is the possibility to set up a synonyms file. So, for example, if your in-house structures use multiple naming conventions for water (*HOH*, *H2O*, *TIP*) these can all be converted to the standard format expected by Relibase+ (*HOH*) using the synonyms file. Similar synonyms can be useful for correcting the naming of other crystallisation reagents, such as glycerol and citrate, which might have been inconsistently named over time.

- Relibase+ also has the capacity to store structure factors with pre-calculated electron density map coefficients, which can be displayed as electron density maps in the web-based viewer. Structure factors can be deposited to Relibase+ along with the corresponding PDB file at the beginning of the data processing.

- Relibase+ data processing has a web-based graphical user interface as well as a command line interface. The web-based graphical user interface makes it easy to upload several structures at a time. The command line interface makes it easy to process a large backlog of structures and to set up automated workflows for getting structures into Relibase+.

3 Ligand templates

- Ligand templates are required to ensure correct atom and bond typing of ligands as the PDB file format does not explicitly contain bond type information.
- However, having to validate templates of all ligands in all in-house structures can easily become cumbersome. The ligand template matching workflow has therefore been designed to minimise the amount of manual intervention required.
- During the data processing the ligands are extracted from the PDB file. The three letter code of the ligand is then used to check if an identically named template already exists in the main Relibase+ database (*reli*). This step is meant to catch ligands that have already had their atom and bond types assigned, which is particularly useful for common crystallisation reagents such as glycerol and citrate.
- If the three letter code does match a template in the *reli* database the template and the ligand are compared using substructure matching. This step filters out any false positives where a ligand in the PDB file has, for some reason or other, been given a three letter code matching a different compound in *reli*.
- If a ligand does not match an entry from the *reli* database the ligand is matched against all user-supplied templates provided during the data input. Matching against the user-supplied templates is also performed using substructure matching.
- If no hit is found processing performs an advanced automatic determination of what the atom and bond typing should be. This *auto-typing* can be automatically accepted by setting the `auto_accept_template` flag to `true` in the relibase_processing.conf file.
- Alternatively, if the `auto_accept_ligand_template` flag is set to `false` the user will be prompted to manually validate the ligand template before it is accepted.



- Note that user-supplied templates should be in mol2 file format.

4 Synonyms file

- A problem that can occur with in-house PDB files is that compounds such as water, citrate and glycerol are given differing three letter codes over the years.
- In terms of water molecules this presents a problem for Relibase+ data processing as it expects them to be marked as *HOH* in the PDB file.
- If compounds such as glycerol and citrate have not been given their official three letter code this means that the ligand template matching algorithm will not be able to automatically assign them a template, thus increasing the amount of manual intervention required. The use of the synonyms file was designed to overcome these types of problems.
- The synonyms file can be used to standardise the use of ligand three letter codes. This is achieved by using a lookup table to allow the substitution of synonyms to a common three letter code.
- Note that when using the synonyms file functionality Relibase+ will not only change the

HETATM records, but also the ATOM (in case of synonyms of modified amino acids), HETNAM, HETSYN, FORMUL, SEQRES, LINK, CISPEP, and MODRES records. Some of these records are not explicitly used by Relibase+ but are modified in order to produce valid PDB files when exporting structures from Relibase+. No effort is made to alter three letter codes in REMARK records.

- The synonyms file can be found in $RELIBASE_ROOT/processing/synonyms.txt.
- To make H2O, D2O and TIP synonyms of HOH add the line:

  HOH H2O D2O TIP

- The first three letter code is the code that the subsequent three letter codes will be converted into.

5 Customising the processing requirements

- Many in-house PDB files do not strictly adhere to the PDB file format, in particular they may not contain all the PDB records Relibase+ expects them to have. In these cases the user will be prompted to add the required field to the PBD file. However, this can become cumbersome if all of the in-house structures are lacking that particular record. Relibase+ therefore allows the user to define the strictness of the data processing, in terms of which fields are required and the syntax allowed in the PDB file. This customisation is defined in the relibase_processing.conf file located in the $RELIBASE_ROOT/processing directory.
  - header_required=true|false
    This flag determines whether the HEADER record is required or not.
  - title_required=true|false
    This flag determines whether the TITLE record is required or not.
  - compound_required=true|false
    This flag determines whether the COMPND record is required or not.
  - source_required=true|false
    This flag determines whether the SOURCE record is required or not.
  - method_required=true|false
    This flag determines whether the EXPDTA record is required or not.
  - author_required=true|false
    This flag determines whether the AUTHOR record is required or not.
  - reference_required=true|false
    This flag determines whether the JRNL record is required or not. (Relibase+ only parses the TITL and REF sub-records of the JRNL records.)
  - crystallographic_data_required=true|false
    This flag determines whether the CRYST1 or SCALEn records are required or not.
  - cryst1_z_value_required=true|false

This flag determines whether or not the z-value of the CRYST1 record is required or not. Most refinement packages do not write out the z-value to the PDB files that they produce.

- deposition_date_required_in_file=true|false
  If the deposition date is not in the PDB file, this flag will prompt the user to add a date to the file. If this flag is set to false the date will be set to that when the structure was uploaded into Relibase+.
- always_use_current_date_as_deposition_date=true|false
  This flag determines if today's date (if the flag is set to true) or the date from the PDB HEADER (if the flag is set to false) should be used as the deposition date.
- negative_residue_numbers_permitted=true|false
  This flag determines whether or not negative residue numbers are permitted.
- synonym_file=/path/to/file
  This flag allows users to specify alternate ligand three letter code synonym files. Relative paths are set to $RELIBASE_ROOT, so if you had a file called alt_synonyms.txt in the $RELIBASE_ROOT/processing directory you could specify it using the flag: residue_name_synonym_file=processing/alt_synonyms.txt.
- auto_accept_ligand_templates=true|false
  This flag can be used to automatically accept the ligand atom and bond typing assigned by Relibase+.
- match_against_pdb_templates=true|false
  If this flag is set to true Relibase+ will attempt to use ligand templates from the main Relibase+ database as input templates based upon matching the ligand three letter code and substructure matching.
- repair_incomplete_conect_records=true|false
  The default option for this flag is set to false. In this case Relibase+ will only try to guess which atoms are connected to each other (a separate concept from determining atom and bond types) if there are no CONECT records present. If there are CONECT records present these will be used to determine which atoms are connected to each other. However if a PDB file, for some reason or other, contain ligands that only have part of their connectivity explicitly defined by CONECT records Relibase+ will not identify these bonds. This can be overcome by setting the repair_incomplete_conect_records flag to true. However, bear in mind that this can lead to non-bonded atoms becoming incorrectly connected to each other if they have coordinates that are anomalously close to each other.
- pdb_templates_exceptions_file=filename
  This option can be used to specify ligand three letter codes that should always be matched against templates in the main Relibase+ database, even if the match_against_templates option is set to false. Simply list the three letter codes of interest on new lines in the file. For example:

  ```
  GOL
  ```

```
ICA
```

Alternatively, if the `match_against_pdb_templates` option is set to true the exceptions file can be used to list three letter codes that should never be matched (for example, `LIG`, `INH`, `UNK`). To list such exceptions prefix the three letter code with an exclamation mark, for example:

```
!LIG
!INH
!UNK
```

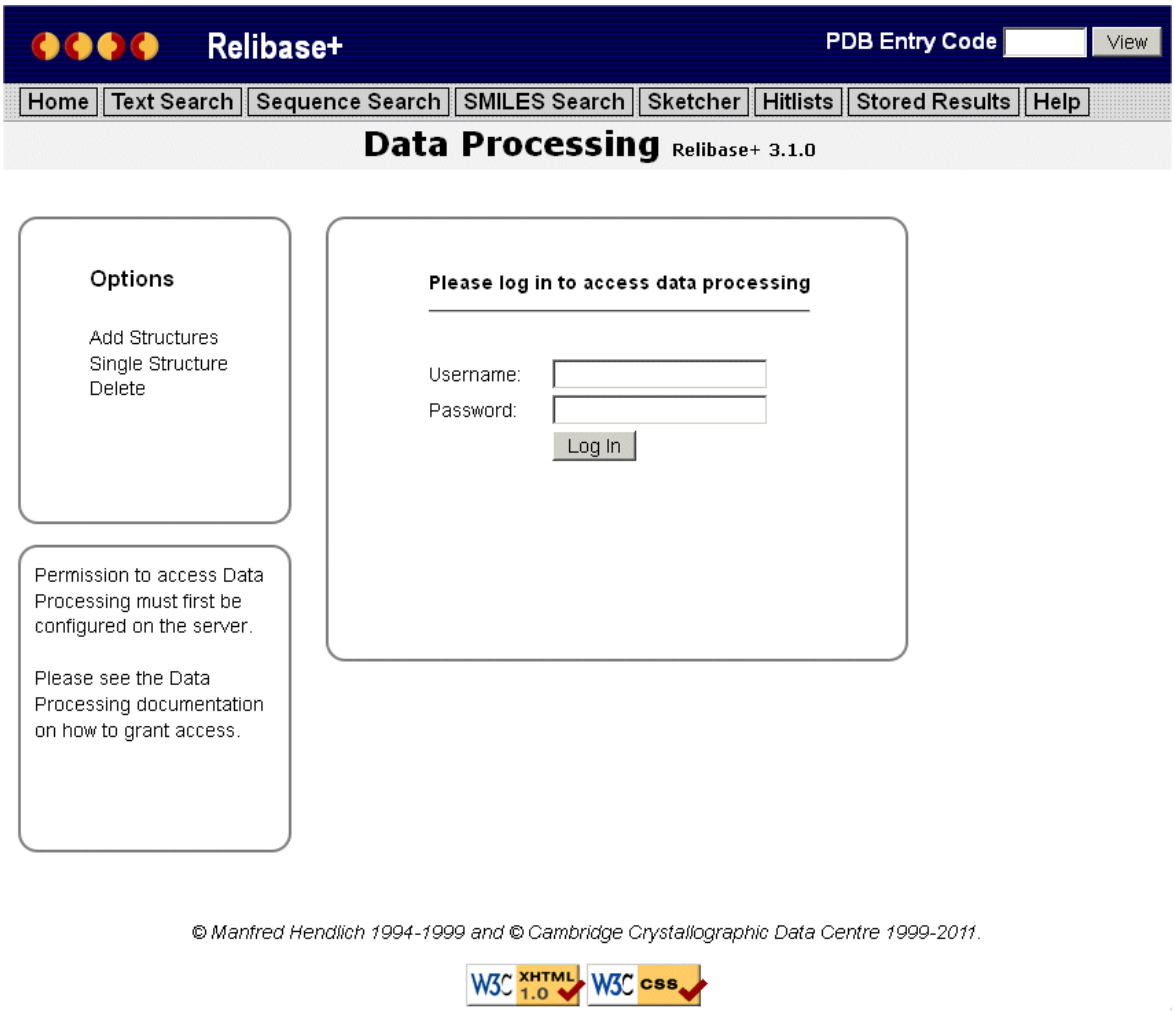
Note that relative paths will be assumed to start from `$RELIBASE_ROOT`, so to specify the file `$RELIBASE_ROOT/processing/exceptions.txt` you would use the line:

```
pdb_templates_exceptions_file=processing/excepitions.txt
```

6 Structure factors and electron densities

- Relibase+ can be used to store structure factors with pre-calculated map coefficients (for both 2Fo-Fc and Fo-Fc maps) and to display electron density maps. In order to store the structure factors MTZ files can be uploaded along with the associated PDB at the data input stage of the data processing.
- The structure factors need to be in MTZ file format. In order for Relibase+ to parse MTZ files it needs to know the column names of the map coefficient data. These typically depend on the software used to generate the MTZ file. At the moment Relibase+ can handle files from the following third party software:
  - Refmac, *http://www.ccp4.ac.uk/html/refmac5.html*, (column names: "FWT", "PHWT", "DELFWT", "PHDELWT")
  - autoBUSTER, *http://www.globalphasing.com/*, (column names: "2FOFCWT", "PH2FOFCWT", "FOFCWT", "PHFOFCWT")
  - Phenix, *http://www.phenix-online.org/*, both with and without fill (column names: "2FOFCWT"/"2FOFCWT_no_fill", "PH2FOFCWT"/"PH2FOFCWT_no_fill", "FOFCWT", "PHFOFCWT").
- If you have MTZ files that use different column names please get in contact so that we can add the capability to parse them.
- There are no dependencies for uploading and storing MTZ files. However, in order to view electron densities in the web-based visualiser, CCP4 (*http://www.ccp4.ac.uk/*) software and libraries are required (see APPENDIX C: Electron Density Configuration and Viewing Section , page 183).

- To upload or delete structures from your in-house database(s) click on the **In-house Database Building Tool** hyperlink on the Relibase+ home page. This will prompt you for a username and password.



- It is necessary to obtain permissions to upload and delete structures (see Obtaining permissions to upload and delete structures Section 7.1, page 164).

7.1 Obtaining permissions to upload and delete structures

- To register a user for permissions to upload and delete structures you will have to run the command below (in a shell that has sourced the `$RELIBASE_ROOT/bin/relibase.setup.sh` on the server) (see APPENDIX E: The Master Relibase+ Command Section , page 187):

```
$ relibase -dpg_user_register <username> <password>
```

- To find out the names of registered workspaces run the `relibase - workspace_editor` commands.
- The user to be registered must already have a workspace. To check whether a user has privileges to upload and delete structures you can use the command:

  ```
  $ relibase -dpg_user_check <username>
  ```

- Note that there are also commands for removing processing priviliges from a user

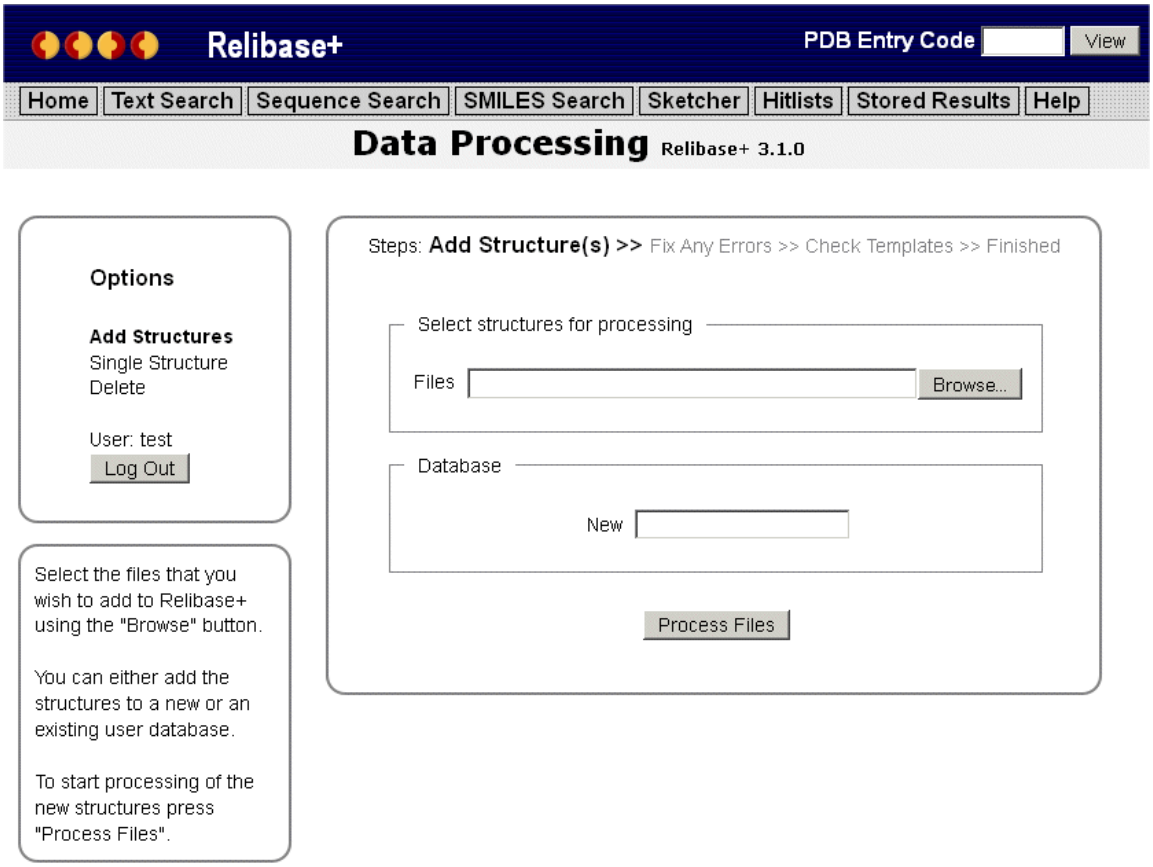
  ```
  $ relibase -dpg_user_delete <username>
  ```

  and for printing a xml list of users that have data processing privileges:

  ```
  $ relibase -dpg_user_print
  ```

7.2Adding a structure

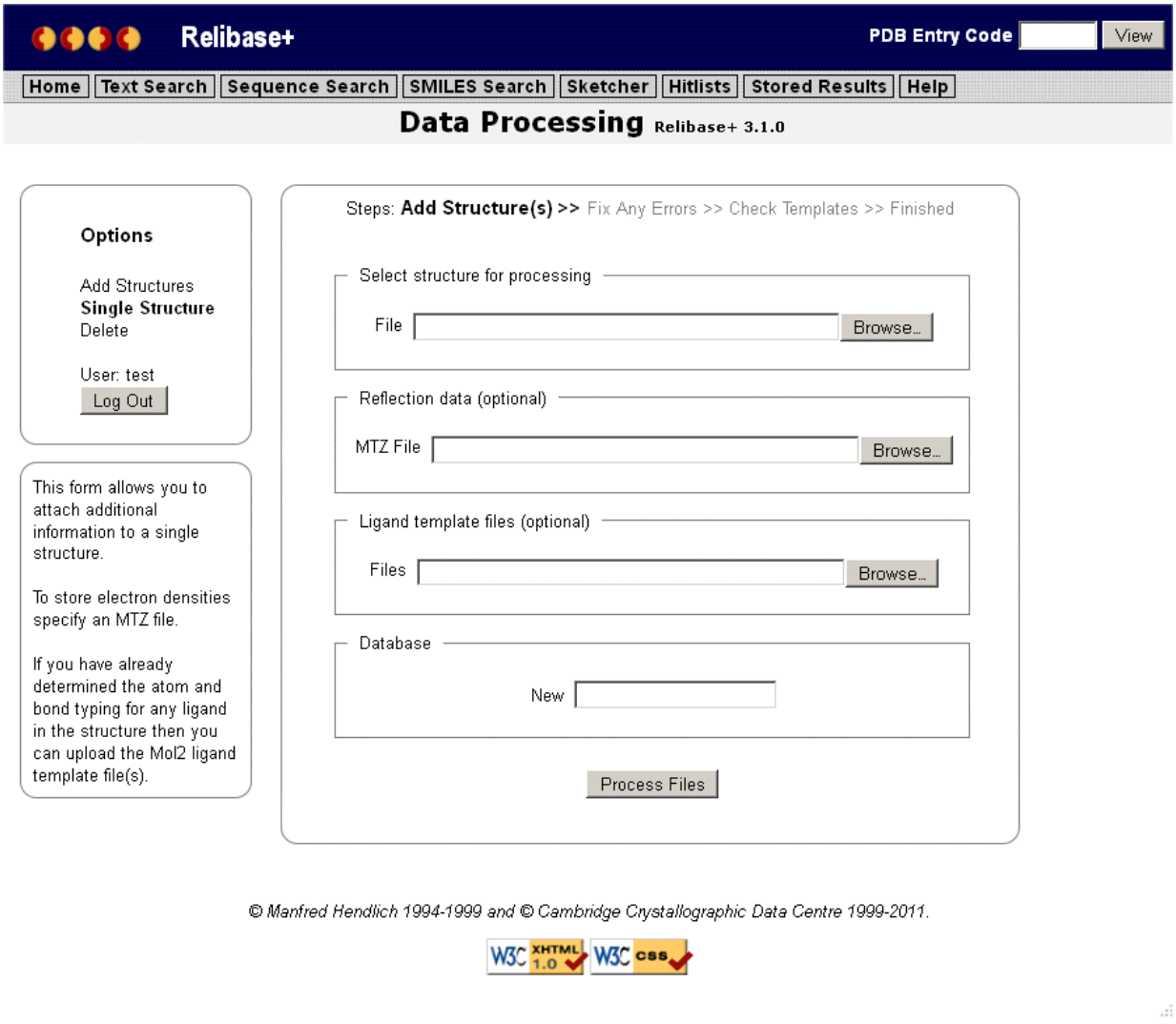
- To add one or more structures to Relibase+ click on the **Add Structures** hyperlink on the left hand side of the data processing page and follow the instructions provided in the help box.



© Manfred Hendlich 1994-1999 and © Cambridge Crystallographic Data Centre 1999-2011.

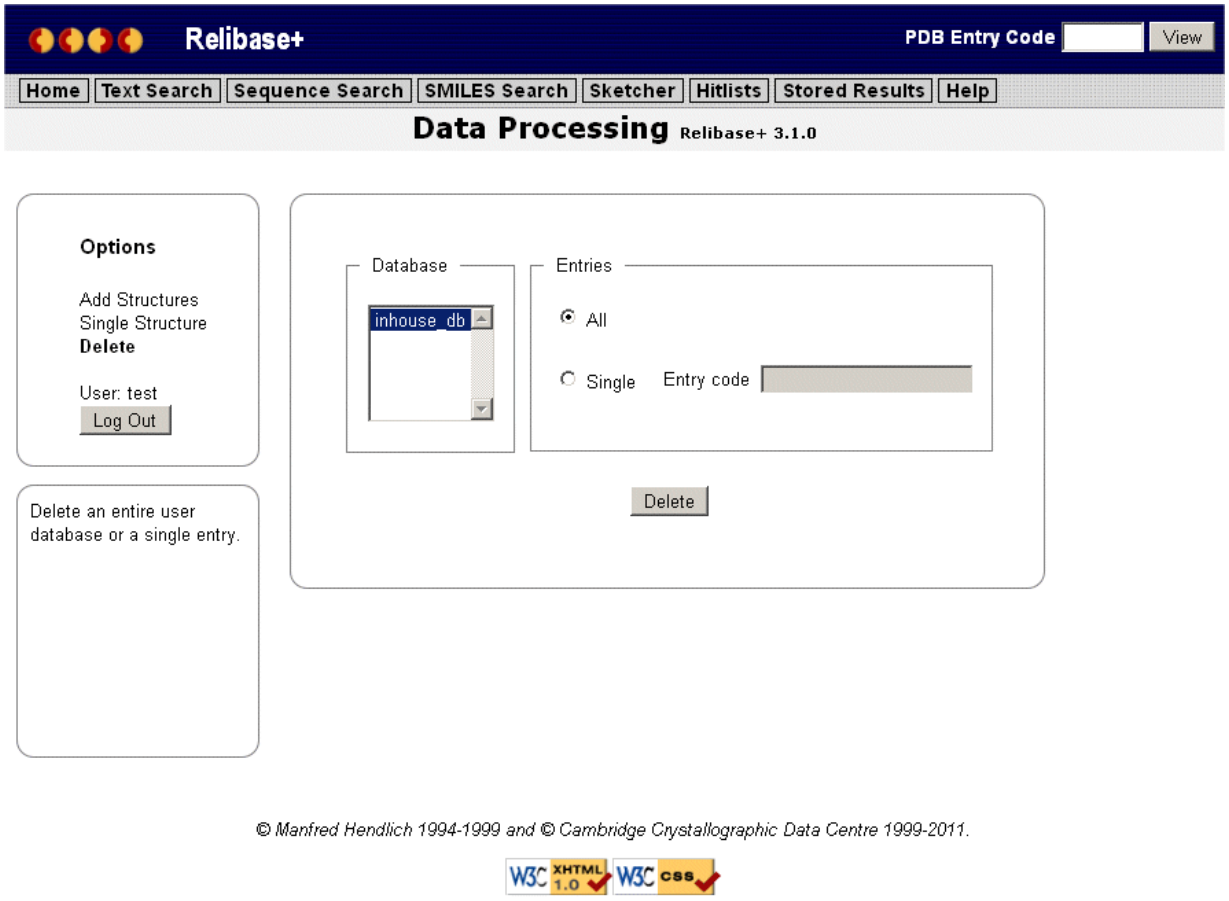
7.3 Adding a structure with associated MTZ and/or ligand templates

- To add a PDB file with associated MTZ and/or ligand templates click on the **Single Structure** hyperlink on the left hand side of the data processing page and follow the instructions provided in the help box.



7.4Deleting a structure or a database

- To delete a structure or a database click on the **Delete hyperlink** on the left hand side of the data processing page and follow the instructions provided in the help box. Note that to delete a single entry you will need to provide the full name of that entry (the full name is taken from the name of the PDB file uploaded to Relibase+).



- Using the settings above, all entries in the database *inhouse_db* would be deleted when the **Delete** button is clicked on.
- If only a single entry was required to be deleted, e.g. *inhouse_structure1.pdb*, then the **Single** radio button should be activated indicating only one entry is being deleted. The filename of the entry you wish to delete should then be entered into the *Entry code* box (i.e. *inhouse_structure1.pdb*). This entry will be deleted from the database when the **Delete** button is selected.

7.5Cancelling work in progress

- If halfway through processing your structures you, for some reason or other, want to cancel the data processing this can be achieved by clicking on the **Cancel** hyperlink on the left hand side of the data processing page.

8 Processing structures using the command line

- Structures can be added to Relibase+ using the command line option below. Further information is available (see APPENDIX E: The Master Relibase+ Command Section , page 187):

```
$ relibase -data_process   input=filename_or_dir
                           database=specification
                                 [conf=configuration_file]
                                 [mtz=mtz_file]
                                 [template=template_file]
```

8.1Processing a single entry

- Suppose that you had the files `2BYH.pdb`, `2BYH_2D7.mol2` (the ligand template file) and `2byh_sigma.mtz` and you wanted to add the entry to a database named `aaa` then you could simply use the command:

  ```
  $ relibase -data_process input=/path/to/dir/2BYH.pdb database=aaa
  mtz=/path/to/dir/2byh_sigma.mtz template=/path/to/dir/2BYH_2D7.mol2
  ```

- This would add your structure to the database.
- Note that the MTZ and template flags are optional. However, if you do not provide a ligand template the structure will not be entered into the database if that template needs user validation. Instead you will get a message stating that:

  ```
  There are ligand templates that require confirmation before this entry
  can be added to Relibase+.
  Please review the files in:
  $RELIBASE_ROOT/processing/dp_cmd/PRE_TEMPL
  Edit if required, then copy to:
  $RELIBASE_ROOT/processing/TEMPL
  Then re-process your input file.
  ```

- At this stage you will have to manually validate the templates in `$RELIBASE_ROOT/processing/dp_cmd/PRE_TEMPL` before copying them to `$RELIBASE_ROOT/processing/TEMPL`.
- To re-process your input file re-run the the first command given above, i.e.

  ```
  $ relibase -data_process input=/path/to/dir/2BYH.pdb database=aaa
  mtz=/path/to/dir/2byh_sigma.mtz template=/path/to/dir/2BYH_2D7.mol2
  ```

8.2Processing multiple PDB files in a directory

- It is possible to provide a directory with PDB files to process. However, we would discourage the use of this option as it is more cumbersome. Suppose that you had some structures in `/home/olsson/structures/` and you wanted to add them to the database `aaa`. The command below would be used to start the process.

```
$ relibase -data_process input=/home/olsson/structures/ database=aaa
```

- Note that at this stage it is highly likely that some of your structures will not have been added to the database because Relibase+ wants you to check your ligand templates. These are located in the `$RELIBASE_ROOT/processing/dp_cmd/PRE_TEMPL` directory unless validation is switched off.
- After you have validated your ligand templates you need to move them from the `$RELIBASE_ROOT/processing/dp_cmd/PRE_TEMPL` directory to the `$RELIBASE_ROOT/processing/TEMPL` directory. You can then add your structures to the database using the same command as used previously:

  ```
  $ relibase -data_process input=/home/olsson/structures/ database=aaa
  ```

  Note that if you forget to move or copy the templates from the `$RELIBASE_ROOT/processing/dp_cmd/PRE_TEMPL` directory to the `$RELIBASE_ROOT/processing/TEMPL` directory before you re-run the command no additional structures will be added to the database as the templates need to be in the latter directory to be regarded as accepted.
- If you want to get a set of structures from a particular directory into a Relibase+ in-house database and you do not want to validate the ligand templates this can be achieved by setting the `auto_accept_ligand_template` flag to `true` in the `relibase_processing.conf` file.
- To re-iterate we discourage the use of this option as it is more cumbersome than the single entry command line option. The single entry command line option can be used to upload a large backlog of in-house structures (see Processing a large backlog of in-house structures Section 8.3, page 170).

8.3Processing a large backlog of in-house structures

- Suppose that you want to process a large number of in-house structures. In this case the preferred method is to create a wrapper script around the single entry command line tool. This is particularly effective if you have mol2 files of the ligands associated with the structures that you want to upload. In these cases a script along the lines of the following *Python script* would be ideal.
- If you do have a large number of in-house structures to process and are unsure of how to go about this please do not hesitate to contact *support@ccdc.cam.ac.uk*.

8.4Using an alternative configuration file

- In cases where people have different types of quality or source of PDB files it might make sense to specify a set of configuration files. For example suppose that a molecular modeller wanted to create an in-house database consisting of PDB files derived from a molecular dynamics (MD)

simulation. The MD derived PDB file might be lacking most PDB header records and as such would need a more permissive set of configuration file settings. In this example the modeller might set the `auto_accept_template_ligand` flag to `true` and disable the requirement of all the missing records in the MD PDB file. Suppose these settings were saved in a file called `/path/to/md_processing.conf` and the MD snapshot PDB files were located in a directory called `/path/to/structures`, the modeller could then upload all the files to a database called `md_snapshots` using the command:

```
$ relibase -data_process input=/path/to/structures/
database=md_snapshots conf=/path/to/md_processing.conf
```

Note that Relibase+ has hardcoded defaults for the data processing configuration options; these are overridden by the flags set in the `relibase_processing.conf` file. When using an alternative configuration file the options are set in the following order:

1. The hardcoded defaults get set.
2. The relibase_processing.conf options are parsed and set.
3. Any options set in the alternative configuration file are parsed and set.

8.5Deleting a structure or an in-house database

- In-house structures and databases can be deleted from Relibase+ using the command below:

```
$ relibase -data delete database=specification
[entry=entry_to_be_deleted]
```

- So suppose you wanted to delete a database named `aaa` then you would use the command:

```
$ relibase -data delete database=aaa
```

- Whereas if you wanted to delete an entry named `snapshot1` from a database called `md_snapshots` you would use the command:

```
$ relibase -data delete database=md_snapshots entry=snapshot1
```

References

ReLiBase

- *Databases for Protein-Ligand Complexes*
  M. Hendlich
  Acta Crystallographica, **D54**, 1178-1182, 1998

VODAK

- *Constituting a Receptor-Ligand Information Base from Quality-Enriched Data*
  K. Hemm, K. Aberer, M. Hendlich
  Intelligent Systems for Molecular Biology, 170-179, 1995

BALI

- *BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank*
  M. Hendlich, F. Rippmann, G. Barnickel
  Journal of Chemical Information and Computer Sciences, **37**, 774-778, 1997

Relibase+

- *Use of Relibase for Retrieving Complex 3D Interaction Patterns Including Crystallographic Packing Effects*
  A. Bergner, J. Günther, M. Hendlich, G. Klebe, M. Verdonk
  Biopolymers (Nucleic Acid Sci.), **61**, 99-110, 2002
- *Relibase - Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions*
  M. Hendlich, A. Bergner, J. Günther, G. Klebe
  J. Mol. Biol., **326**, 607-620, 2003
- *Utilising Structural Knowledge in Drug Design Strategies - Applications Using Relibase*
  J.Günther, A. Bergner, M. Hendlich, G. Klebe
  J. Mol. Biol., **326**, 621-636, 2003

Uppsala Electron-Density Server

- *The Uppsala Electron-Density Server*
  G. J. Kleywegt, M. R. Harris, J. Zou, T. C. Taylor, A. Wählby, T. A. Jones
  Acta Cryst., **D60**, 2240-2249, 2004.
  [DOI:*10.1107/S0907444904013253*]
  http://eds.bmc.uu.se/

Water Information Module (WaterBase)

- *Cluster Analysis of Consensus Water Sites in Thrombin and Trypsin Shows Conservation*

Between Serine Proteases and Contributions to Ligand Specificity
P. C. Sanschagrin, L. A. Kuhn
Protein Sci., **7**, 2054-2064, 1998

- *Valence Screening of Water in Protein Crystals Reveals Potential Na+ Binding Sites*
  M.Nayal, E. Di Cera
  J. Mol. Biol., **256**, 228-234, 1996
- *Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions*
  H.Gohlke, M.Hendlich, G.Klebe
  J. Mol. Biol., **295**, 337-356, 2000

Cavity Information Module (CavBase)

- *LIGSITE: Automatic and efficient detection of potential small molecule binding sites in proteins*
  M. Hendlich, F. Rippmann and G. Barnickel.
  J. Mol. Graph. Model., **15**, 359-63, 389, 1997
- *From Structure to Function: A New Approach to Detect Functional Similarity among Proteins Independent from Sequence and Fold Homology*
  S. Schmitt, M. Hendlich and G Klebe
  Angew. Chem. Int. Ed. Engl. **40**, 3141-3144, 2001
- *A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology*
  S. Schmitt, D. Kuhn and G. Klebe
  J. Mol. Biol. **323**, 387-406, 2002
- *Structural Aspects of Binding Site Similarity: A 3D Upgrade for Chemogenomics. In Chemogenomics in Drug Discovery*, (Eds Hugo Kubinyi and Gerhard Müller), Wiley-VCH, Weinheim (2004)
  A. Bergner and J. Günther.

Secondary Structure Module (SecBase)

- *Turns revisited: A uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions.*
  O. Koch and G. Klebe
  Proteins: Structure, Function, and Bioinformatics, 74, 353-367, 2008
  [DOI: *10.1002/prot.22185*]
- *Secbase: Database Module To Retrieve Secondary Structure Elements with Ligand Binding Motifs*
  O. Koch, J. Cole; P. Block, G. Klebe
  J. Chem. Inf. Model., **49**, 2388-2402, 2009
- *Prediction of turn types in protein structure by machine-learning classifiers*
  M. Meissner, O. Koch, G. Klebe, G. Schneider
  Proteins,. **74**, 344-352., 2009

Acknowledgements

Relibase(+) Development

The following people were involved in the development of ReLiBase prior to the CCDC taking over the onward development and maintenance of the program, now known as Relibase+:

- Dr. Manfred Hendlich
- Dr. Gerhard Barnickel
- Dr. Klemens Hemm and Dr. Karl Aberer
- Dr. Ingo Dramburg
- Dr. Judith Günther
- Dr. Stefan Schmitt
- Dr. Andreas Bergner
- Prof. Gerhard Klebe
- Dr. Oliver Koch

Third-Party Software in Relibase+

The following third-party software is used in Relibase+:

- The FASTA package for sequence searching and alignment by Bill Pearson (*http://www.people.Virginia.EDU/~wrp/pearson.html*).
- Dr. Henry Spencer's regex library (Regular Expression Library).
- Code from the SPLASH library by Dr. Jim Morris (*http://www.wolfman.com/splash.html*).
- The ligand 2D diagrams were generated using the CACTVS Toolkit by Dr. Wolf-D. Ihlenfeldt (University of Erlangen, Germany *http://www2.ccc.uni-erlangen.de/software/cactvs/*) and ChemDraw (*http://www.camsoft.com*) and Marvin by ChemAxon Ltd.
- Embedded visualisation powered by AstexViewer™ (*http://www.astex-therapeutics.com/AstexViewer/*).
- AstexViewer™ also incorporates Thinlet (*http://thinlet.sourceforge.net/home.html*).

Other Acknowledgements

- The staff of the Protein Data Bank at Brookhaven National Laboratory, USA, who maintained and developed the PDB archive until 1998.
- The members of the Research Collaboratory for Structural Bioinformatics (RCSB), responsible for the maintenance of the PDB (*http://www.rcsb.org/pdb/*) since 1998.

- Dr. Thomas Mietzner (BASF, Ludwigshafen, Germany) for providing code for the superposition of molecules, based on atom pairs.
- The German Federal Ministry of Education and Research (BMBF), Merck (Darmstadt, Germany), BASF (Ludwigshafen, Germany) and Boehringer Ingelhheim (Germany) for funding the RELIWE and RELIMO projects.
- Dr. Stefan Schmitt in the group of Gerhard Klebe at the Institute of Pharmaceutical Chemistry, Philipps-University Marburg, Germany for the cavity information module (CavBase).
- Dr. Oliver Koch, jointly at CCDC and in the group of Gerhard Klebe at the Institute of Pharmaceutical Chemistry, Philipps-University Marburg, Germany for the secondary structure information module (SecBase).

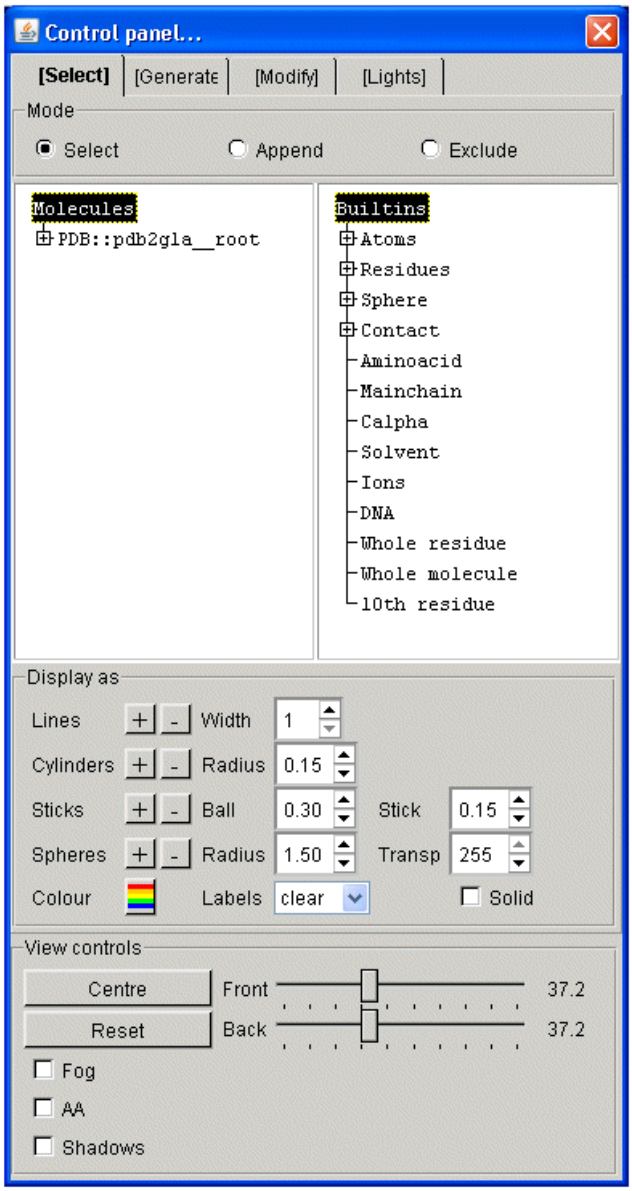
APPENDIX A: Keyboard Shortcuts for the Relibase+ Sketcher

- Keyboard shortcut options that are available for the Relibase+ sketcher are provided below.

Keyboard Shortcut	Function
CTRL-A	Selects everything in the sketcher (see Selecting Atoms Section 2.7, page 123).
CTRL-I	Everything that is selected in the sketcher becomes unselected, and vice versa (see Selecting Atoms Section 2.7, page 123).
CTRL-SHIFT-A	Deselects everything in the sketcher (see Selecting Atoms Section 2.7, page 123).
Delete	Deletes everything that is selected in the sketcher (see Deleting Atoms and Bonds Section 2.8, page 125).
CTRL-C	Copies all selected atoms in the sketcher (see Duplicating Substructures (Copy and Paste) Section 8.4, page 137).

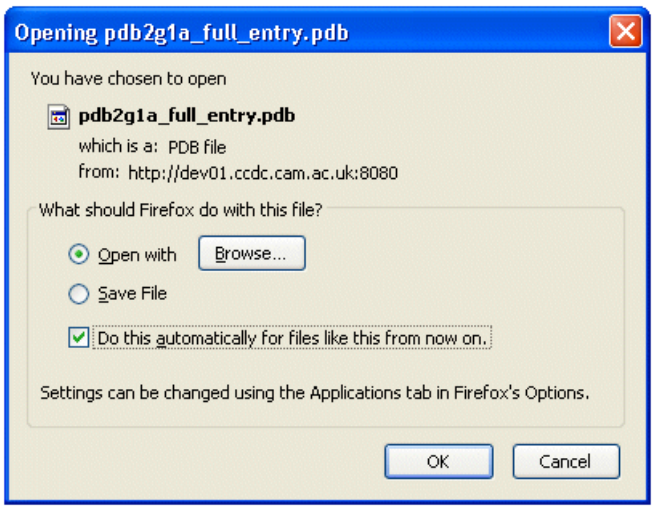
APPENDIX B: Making the Most Out of Visualisation

- The ability to easily and conveniently visualise proteins and protein-ligand complexes is central to the design of Reliabse+. Below are some suggestions on how you can make the visualisation work for you.
- The web-based visualiser is very powerful, many selection and visualisation options can be accessed by right clicking in the visualiser and choosing **Select**, **Popup...** or by pressing the **F11** key on the keyboard.



- The Hermes visualiser has been tailored to work with Relibase+ particularly in terms of making use of the information from the water and the cavity modules of Relibase+. Hermes can be accessed from all protein and ligand pages by clicking the **Show in Hermes** button. However, some users simply want a quick way of loading the protein structures into their favourite

visualiser. For this purpose there is a **Save PDB File** button at the top of the protein pages and a **Save Complex PDB File** button at the top of ligand pages. The behaviour of these download buttons can be customised, in your web-browser, to automatically open the downloaded files in a 3rd party visualiser of preference.



- Finally, note that on top of protein pages that have structure factors associated with them there is a button named **Save PDB+MTZ File**. This can be used to download a tar file containing both the PDB file and the associated MTZ file. By writing a custom script to handle this archive it is possible to set up a system where PDB files are automatically displayed with their associated electron density maps in a visualiser such as Coot by pointing the browser at a script such as the one outlined below:

```
#!/bin/sh
# make sure COOT is in the path
PATH=/where/ever/coot/bin:$PATH

# create temporary directory
mkdir /tmp/$$.dir
cd /tmp/$$.dir

# unpack tarball:
tar -xf $1
pdb=`ls *.pdb | head -n 1`
mtz=`ls *.mtz | head -n 1`

# fire off coot:
coot --pdb $pdb --auto $mtz

# exit and remove temporary directory
```

```
    cd /tmp && rm -fr /tmp/$$.dir
    exit 0
```

Note that if you want the **Save PDB+MTZ File** button to produce a *zip* file instead of a *tar* file this can be configured in the `$RELIBASE_ROOT/relibase_htdocs/include/global_settings.php`. Change the default from *tar* to *zip* in the line:

```
define("PDBMTZ_ARCHIVE_FORMAT", "tar"); // zip or tar
```

APPENDIX C: Electron Density Configuration and Viewing

Electron Density Configuration:

- Relibase+ has the ability to store structure factors and to display electron density maps in the web-based visualiser. However Relibase+ does not have built in software for converting structure factors to electron density maps. It therefore requires access to CCP4 (*http://www.ccp4.ac.uk/*) libraries and software. If Relibase+ does not have access to CCP4 libraries and software it will work as normal, but you will not have the ability to visualise electron density maps in the web-based visualiser.
- During the installation you will be asked to (optionally) specify the location of your CCP4 libraries and software. If you wish to add or alter this path post installation you will need to manually edit your `$RELIBASE_ROOT/bin/relibase.setup.sh` (sh or bash shells) and `$RELIBASE_ROOT/bin/relibase.setup` (tsch shell) files. For the former (sh or bash shells) add the line:

  ```
  . $CCP4_MASTER/setup-scripts/sh/ccp4.setup
  ```

  to the end of the `relibase.setup.sh` file (note that `$CCP4_MASTER` should be the top level of your CCP4 software installation). For example:

  ```
  . /local/ccp4/setup-scripts/sh/ccp4.setup
  ```

  For the latter (tsch shell) add the line:

  ```
  source $CCP4_MASTER/setup-scripts/csh/ccp4.setup
  ```

  to the end of the `relibase.setup` file. For example:

  ```
  source /local/ccp4/setup-scripts/csh/ccp4.setup
  ```

- Structure factors for in-house structures can be stored in Relibase+. For structures in the core *reli* database structure factors can be fetched on a per entry basis from the Uppsala Electron Density Server (EDS) (*http://eds.bmc.uu.se/eds/*) if available. Downloaded MTZ files are cached in `$RELIBASE_ROOT/relibase_htdocs/tmp/EDS_MTZ`. To disable caching of EDS MTZ files set the `DISABLE_EDS_CACHE` parameter to `1` (`0` to enable it) in the `$RELIBASE_ROOT/bin/relibase.setup.config` file. To disable access to the EDS server completely set the `DISABLE_EDS_ACCESS` parameter to `1` (`0` to enable it) in the `$RELIBASE_ROOT/bin/relibase.setup.config` file.
- If you are using a proxy server, Relibase+ will additionally need to be configured with your proxy server settings in order to communicate with the EDS server. To do so, edit the `JAVA_OPTS=` line in your `$RELIBASE_ROOT/bin/relibase.setup.config` file. You

will need to ensure that the following settings are present:

```
-Dhttp.proxyHost=my.proxy.com -Dhttp.proxyPort=1234
```

where *my.proxy.com* and *1234* are your actual proxy server and port number. For example:

```
JAVA_OPTS="-Xmx512m -Xms256m -verbose:gc -
Dhttp.proxyHost=my.proxy.com -Dhttp.proxyPort=1234"
```

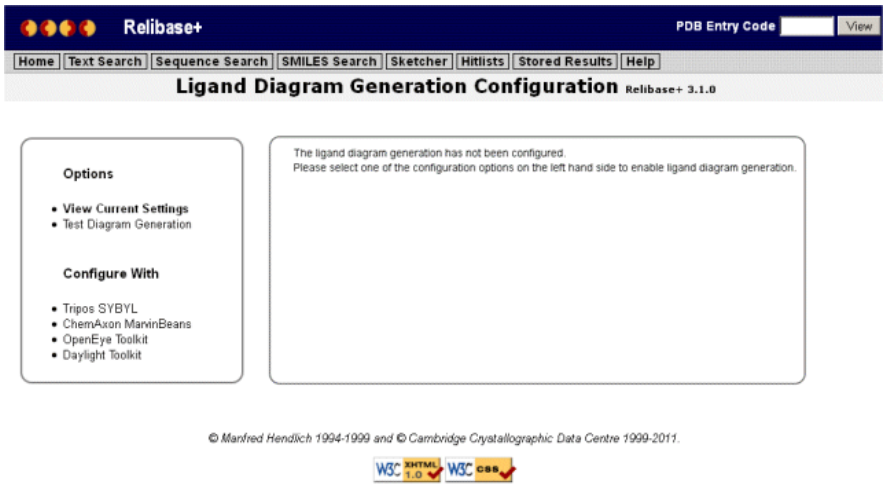
- Note that to make any changes to the relibase.setup.config file take effect you will need to stop your relibase server, source your Relibase setup and restart the relibase server.

Electron Density Viewing:

- Viewing the electron density maps can give an immediate appreciation of the quality of the experimental data upon which a structure is based.
- Electron density maps can be displayed in the web-based visualiser by clicking on the **Load Maps** button. By default both the *2FoFc* (blue) and the *Fo-Fc* (+ve green, -ve red) maps are displayed. If you want to hide any of the maps deselect the relevant checkboxes in the controller.
- The result of a successful X-ray crystal structure determination is the 3-dimensional distribution, or density map, of the electrons in the crystal. However, as the electrons are closely associated with the atomic nuclei, the electron density map, viewed as a contour surface, gives a good 3D representation of the shape and internal structure of the molecules (dependent on both the resolution and degree of error in the experimental diffraction measurements and their processing) and is used to build the atomic model of the structure. A number of forms of the electron density map can be calculated, to highlight different types of information. The two most commonly displayed are the *2Fo-Fc* and *Fo-Fc* maps.
- The *2Fo-Fc* map (blue) provides a representation of how well the model agrees with the experimentally derived electron density distribution. The *Fo-Fc* difference map specifically highlights areas of disagreement between the two. Positive (green) density indicates regions where structure is present but has not been modelled and negative (red) density indicates regions where atoms have been modelled that are not supported by the underlying experimental data. In both cases, the detail visible in the maps will be dependent on the resolution of the experimental data and may contain artefacts due to errors in measurement or processing of the data.

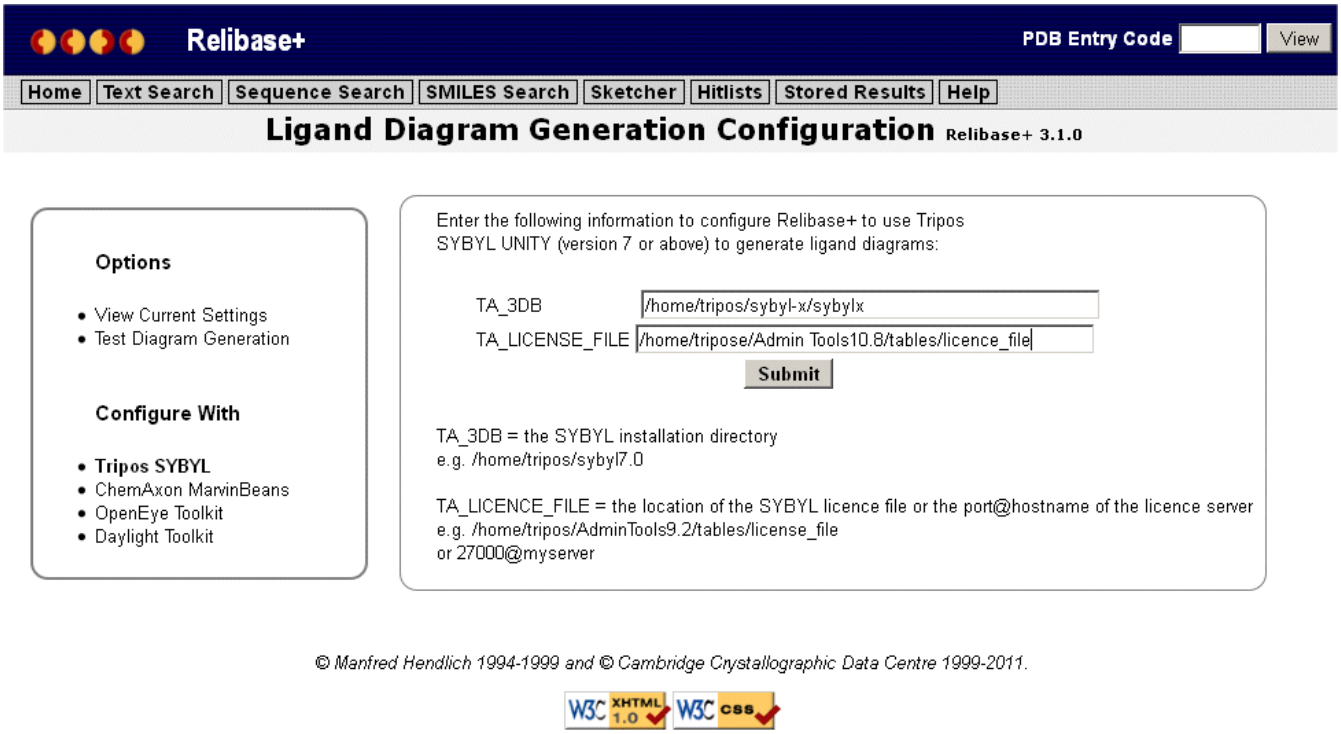
APPENDIX D: Configuring the Ligand Diagram Generation

- In order to generate 2D diagrams for in-house structures you will need to use 3rd party software. Relibase+ can make use of:
  - Tripos SYBYL.
    Note that versions of SYBYL prior to 7.0 are not compatible.
  - ChemAxon MarvinBeans.
    Please note that OpenBabel is also required to convert file formats.
  - OpenEye Toolkit*.
  - Daylight Toolkit*.
    ** Note that the latter two options are experimental and have not been extensively tested in-house.*
- To configure the ligand diagram generation go to the *Help* page and click on the *Ligand Diagram Generation Configuration* hyperlink. This will take you to the *Ligand Diagram Generation Configuration* page.



- Select the package that you want to use for your ligand diagram generation on the left hand side and enter the information required:
  - Tripos SYBYL: path to the SYBYL installation directory (`TA_3DB`) and the location of the licence file or the `port@hostnmae` of the licence server (`TA_LICENCE_FILE`).
  - ChemAxon MarvinBeans: the location of the MarvinBeans converter (`MARVINBEANS_PATH`) and the location of the Babel executable (`BABEL_PATH`).
  - OpenEye Toolkit: the location of the mol2gif converter (`MOL2GIF_PATH`) and the location of the `oe_license.txt` file (`OE_LICENSE`).
  - Daylight Toolkit: the path to the Daylight installation directory (`DY_DIR`) and the location of the Daylight licence file (`DY_LICENCEDATA`).
- Note that the paths and files specified need to be accessible by the machine that the Relibase+

system is installed on.



APPENDIX E: The Master Relibase+ Command

If the Relibase+ environment is set (see the installation guide for further information), the command relibase is aliased to:

```
<RELIBASE_ROOT>/bin/relibase_master.com
```

The following actions and options are allowed:

Relibase+ Server Options:

relibase	**-all**	**start**		
		Starts all Relibase+ servers.		
relibase	**-all**	**stop**		
		Stops all Relibase+ servers		
relibase	**-httpd**	**start**		
		Starts the Relibase+ Apache HTTPD server.		
relibase	**-httpd**	**stop**		
relibase	**-httpd**	**shutdown**		
		Stops the Relibase+ Apache HTTPD server.		
relibase	**-database**	**start**		
		Starts the Relibase+ Derby database server.		
relibase	**-database**	**stop**		
relibase	**-database**	**shutdown**		
		Stops the Relibase+ Derby database server		
relibase	**-database**	**status**		
		Reports the current status of the Relibase+ Derby database server.		
relibase	**-server**	**start**	**[force]**	

		Starts the Relibase+ Derby database server. Note the Derby database must be running before this command can be used. The optional `force` *command will force the Reliabse+ server to start even if it cannot detecet a running database server.*		
relibase	**-server**	**stop**		
relibase	**-server**	**shutdown**		
		Stops the Relibase+ Derby database server.		
relibase	**-server**	**status**		
		Details the current status of the Relibase+ server.		
relibase	**-software**	**update**		
		Attempts to obtain and install the latest software update from the CCDC's ftp server.		
relibase	**-software**	**update**	**package=path_to_file**	
		Updates Relibase+ with the software update package specified by the path given in the `package=` *argument.*		

Relibase+ Data Commands:

relibase	**-data**	**update**		
		Attempts to obtain and install the latest data update from the CCDC's ftp server.		
relibase	**-data**	**update**	**package=path_to_file**	
		Updates Relibase+ with the data update package specified by the path given in the `package=` *argument.*		
relibase	**-data**	**retry_failed**		
		If any entries failed to be applied to the Relibase+ database during a data update, this command will attempt to re-process them.		

relibase	**-data**	**delete**	**database=<db>**	
		Deletes from Relibase+ the entire database <db> *using the* database= *argument (see Deleting a structure or a database Section 7.4*, page 167).		
relibase	**-data**	**delete**	**database=<db>**	**[entry=entry_to_be_deleted]**
		Deletes only the structure specified using the entry= *argument from the database* <db> *using the* database= *argument (see Deleting a structure or a database Section 7.4*, page 167).		
relibase	**-data_process**	**input=filename**	**database=<db>**	**[conf=conf_file]**
				[mtz=mtz_file]
				[template=template_file]
				[force]
	Processes the single file filename *and attempts to add it to the database* <db>*. There are several optional arguments for this command:* conf= *allows you to specify a specific configuration file instead of the default.* mtz= *allows you to specify a MTZ file to be made available from the database entry.* template= *allows you to specify a template mol2 file to be used for a ligand in your input file. This argument may be used multiple times, once for each template being provided.* force *allows you to force entry and ignore any warnings or errors.*			

Relibase+ Export/Import Commands:

relibase	**-dump_ligands**	**database=<db>**	**[radius=float]**	
			[only_ref=on|off]	
			[only_names=on|off]	

				[format=mol2|sdf]
	relibase - dump_ligands database=<db> *writes out all the ligands in the database* <db>*. There are several optional arguments for this command:* radius=float *By default, the radius is zero which means that only ligands will be written to the output files. If set to a larger value (in Angstrom), all protein residues within this distance of the ligand will also be written to the output files.* only_ref=on *writes out all the reference ligands e.g. if there is >1 occurrence of an SO_4 anion, only one ligand will be written out.* only_ref=off *writes out all occurrences of ligands etc in the specified database.* only_names= *This command provides a listing of reference ligands and corresponding ligand models. Used in conjunction with the* only_ref=on *option above, this provides a list of unique reference ligands for which images are required. Note that this command applies to a single database..* format= *Use this command to stipulate which format your ligand file is written out in (i.e. either mol2 or sdf).*			
relibase	**-dump_cavities**	**database=<db>**		
		Exports all cavities, in XML format, from the database <db>*.*		
relibase	**-dump_waters**	**database=<db>**		
		Exports all water information, in XML format, from the database <db>*.*		
relibase	**-dump_pdb_xml**	**database=<db>**		
		Exports all PDB data, in XML format, from the database <db>*.*		
relibase	**-hitlist_upload**	**hitlist_xml_file**		
		Imports the XML-based hitlist file hitlist_xml_file*.*		

Relibase+ User/Administrator Commands@

relibase	**-licence_info**			
relibase	**-check_licence**			
		Details information about your currently Relibase+ licence.		
relibase	**-workspace_editor**			
		Launches the Relibase+ workspace editor to allow administration of user workspaces.This command will launch a java applet that will allow you to delete client workspaces.		
relibase	**-dpg_user_register**	**username**	**password**	
		Enables an existing client workspace `username` *to access the Data Processing GUI with the specified* `password`.		
relibase	**-dpg_user_check**	**username**		
		Displays the current details for the workspace `username`.		
relibase	**-dpg_user_delete**	**username**		
		Removes the Data Processing GUI rights for the specified `username`. *Note that the workspace username will still exist - use the workspace editor to remove it.*		
relibase	**dpg_user_print**			
		Displays information for all users.		

APPENDIX F: Calculating Descriptors: Computational Details

Acceptor Atoms (see page 193)
Atom Types (see page 193)
Buried Atoms (see page 194)
Donatable (or Polar) Hydrogens (see page 194)
Hydrogen Bonds (see page 194)
Hydrophobic Atoms (see page 194)
Number of Buried Hydrogens/Acceptors Not Forming an H-Bond (see page 195)
Polar, non-Hydrogen-Bonding Atoms (see page 195)
Rotatable Bonds (see page 195)
Solvent Accessibility (see page 195)
Solvent Inaccessible Ligand Surface Area (see page 195)
Sphere Accessible Volume (see page 196)

Acceptor Atoms

An acceptor atom is any atom capable of accepting a hydrogen bond. This includes almost all oxygen and nitrogen atoms, but with the exceptions of trigonal planar (R_3N) and quaternary (R_4N) nitrogens, since these have no free lone pairs. Oxygen atoms in conjugated environments (e.g. furan oxygen) and the central nitrogen of an azide group are counted as acceptors, though this is questionable.

Atom Types

Some options in the Relibase+ visualiser allow specification of Sybyl atom types, devised by Tripos Inc. (http://www.tripos.com). Amongst the most common types are:

H	hydrogen
C.1	*sp* carbon
C.2	sp^2 carbon
C.3	tetrahedral carbon
C.ar	aromatic carbon
N.1	*sp* nitrogen
N.2	two-coordinate, non-linear nitrogen
N.3	pyramidal trigonal nitrogen
N.pl3	planar, trigonal nitrogen
N.am	amide nitrogen

N.ar	aromatic nitrogen
O.2	carbonyl oxygen
O.3	ether oxygen
O.co2	carboxylate/phosphate oxygen
S.2	thiocarbonyl sulphur
S.3	thioether sulphur
S.o	sulphoxide sulphur
S.o2	sulphone sulphur
P.3	phosphate phosphorus
Halogens	standard element symbols e.g. F, Cl, Br, I
Metals	standard element symbols, e.g. Zn, Fe, Ca

Buried Atoms

An atom is counted as buried if no part of it is solvent accessible (see Solvent Accessibility, page 195). It does not matter what types of atoms render the atom inaccessible, i.e. whether the atom is buried by protein or ligand atoms or a combination of both.

Donatable (or Polar) Hydrogens

A donatable or polar hydrogen is any hydrogen that can be donated in a hydrogen bond. This includes any oxygen-, nitrogen- or sulphur-bound H atom.

Hydrogen Bonds

By default, a protein-ligand hydrogen bond is present if:

- it involves a recognised polar hydrogen (see Donatable (or Polar) Hydrogens, page 194) and a recognised acceptor atom (see Acceptor Atoms, page 193); and
- the H...acceptor distance is less than the sum of van der Waals radii and the donor-H...acceptor angle is greater than 90.

These criteria can be customised when setting up H-bond descriptors (please refer to the relevant section of the Hermes documentation for further information). Intramolecular hydrogen bonds, either in the protein or the ligand, are not recognised.

Hydrophobic Atoms

The definition of hydrophobic atoms includes almost all types of carbon atom (but not cyanide or carbonyl carbon), sp^3 hybridised sulphur, non-ionised chlorine, bromine and iodine, and any

hydrogen atom that is covalently bonded to a hydrophobic atom (effectively, C-H and S-H). Note that S-H is also counted as a donatable hydrogen.

Number of Buried Hydrogens/Acceptors Not Forming an H-Bond

The following descriptors:

- *Number of buried donatable ligand hydrogen atoms not forming an H-bond*
- *Number of buried ligand acceptors not forming an H-bond*
- *Number of buried donatable protein hydrogens not forming an H-bond*
- *Number of buried protein acceptors not forming an H-bond*

are counts of "occluded" hydrogen-bonding atoms, i.e. atoms that are inherently capable of participating in hydrogen bonds but are prevented from doing so because they are buried by non-H-bonding atoms. For example, a histidine-ring NH would be counted as occluded if it were solvent accessible before docking but was buried by ligand carbon atoms after docking. A limitation of these descriptors is that no allowance is made for intramolecular hydrogen bonding (i.e. in the example just quoted, the NH would still be counted as occluded even if it were H-bonded to a neighbouring protein atom).

Polar, non-Hydrogen-Bonding Atoms

Polar atoms in descriptors such as *Number of buried polar atoms in the ligand* are atoms that have some polar character but cannot participate in hydrogen bonds. The definition of these includes any nitrogen that is not counted as an H-bond acceptor (notably planar-trigonal and quaternary N) and fluorine, sulphur and phosphorus.

Rotatable Bonds

A ligand bond is considered rotatable if it is single, acyclic and not to a terminal atom. This therefore includes, e.g., bonds to methyl groups but not to chloro substituents. It also includes bonds which, although single and acyclic, have highly restricted rotation, e.g. ester linkages. Finally, it incorrectly include bonds to linear groups, e.g. the bond between the methyl and cyanide carbons in CH_3-CN.

Solvent Accessibility

An atom is counted as being solvent accessible if any part of it is exposed to solvent. Solvent accessibility is measured assuming a default solvent radius of 1.4Å, however this value can be altered.

Solvent Inaccessible Ligand Surface Area

This is a measure of how much of the ligand surface area is desolvated upon docking, calculated assuming a solvent molecule radius of 1.4Å (this value is the default value which can be altered). The

surface area of the undocked ligand *in the same conformation* is also written out. Values can be output in units of A^2 or as percentages. If percentage quantities have been requested, the undocked value will always be zero.

Sphere Accessible Volume

Two values are output, before and after docking. The former (undocked) value measures how much of the sphere volume is vacant (i.e. unoccupied by protein atoms) before the ligand is docked. The latter measures how much sphere volume remains vacant (i.e. unoccupied by either protein or ligand atoms) after docking. Volumes may be output in A^2 or as percentages. If percentage quantities have been requested, the undocked value will always be 100 (since, by definition, 100% of the sphere volume not occupied by protein atoms must be vacant before the ligand is docked). If absolute (A^2) values have been requested, the undocked value may vary a little from one docking to another because GOLD may move protein polar H-atoms during docking, therefore altering the amount of sphere volume that is occupied by protein atoms.

APPENDIX G: Tutorials

1 **Tutorial 1: Introduction to the Relibase+ Graphical User Interface** (see page 3)
2 **Tutorial 2: Substructure Searching in Relibase+ (SMILES and 2D/3D)** (see page 21)
3 **Tutorial 3: An Introduction to the Cavity Information Module and Hermes** (see page 31)
4 **Tutorial 4: Using the Cavity Information Module in a More In-depth Way** (see page 43)
5 **Tutorial 5: Introduction to the Secondary Structure Module** (see page 59)

1 Tutorial 1: Introduction to the Relibase+ Graphical User Interface

1.1 Introduction

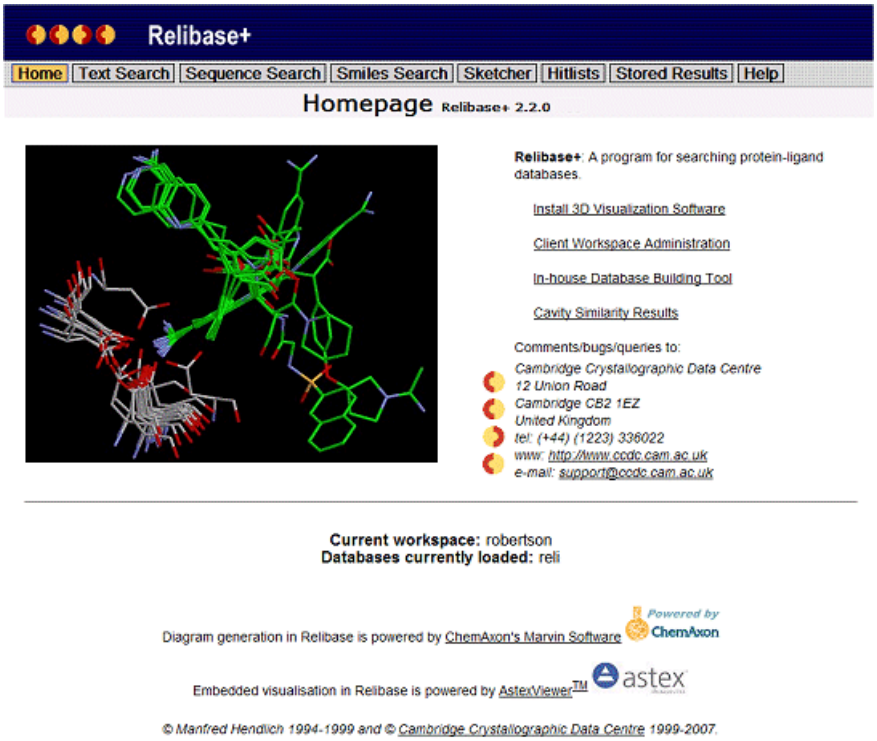
- In order to familiarise yourself with the Relibase+ graphical user interface (GUI) it is recommended that you work through the examples provided. The first few tasks have been designed to introduce the basic types of information present in Relibase+, and the interaction of the GUI with the visualiser used in Relibase+, Hermes.

1.2 Coverage of the Relibase+ Database

- The database behind Relibase+ is the Protein Data Bank (PDB) (*http://www.rcsb.org/pdb/*). It covers all entries in the PDB with the exception of theoretical structures. However, structures where a ligand (substrate) molecule was modelled into an experimental protein structure, *are* included. You can also use Relibase+ to search your inhouse database alongside the PDB.
- It is important to note that in Relibase+, all non-protein moieties in a structure are considered to be either ligands or water molecules. Hence metal ions, anions, solvate molecules, cofactors and inhibitors are all regarded as ligands. In the 3D visualiser, DNA and RNA strands are displayed as ligands, but they are ignored in ligand-substructure searches.
- Each protein entry in the Relibase+ database corresponds to an entry in the PDB and contains the following information:
  - Bibliographic, textual and numerical information
  - Crystal structure data (for X-ray structures)
  - Protein chain(s)
  - Binding site(s)
  - Chemical diagram of the ligand(s)
  - Crystal packing of the protein-ligand binding site
  - Information about the water content of a particular protein
  - Information about any cavities present within a particular protein

1.3 The Relibase+ Home Page

Relibase+ is a web-based application and all its functionality is accessible via a web browser such as Netscape or Internet Explorer. Open Relibase+ from within your browser:



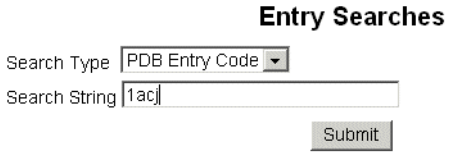
- The 3D visualisation software will already have been installed for you. Two visualisers are provided with Relibase+ and serve slightly different purposes:
  - AstexViewer, which is embedded in the Relibase+ interface to provide quick and easy visualisation of hit structures.
  - Hermes, to facilitate more detailed investigation of hit structures.
- The workspace username, *robertson* is given and the databases that are currently loaded for searching alongside each other. In this case there are three inhouse databases, *battletwo*, *ian* and *jase* in addition to the PDB itself, *reli*.
- From this page in the interface, it is possible to access the Data Processing Graphical User Interface (GUI) for generating inhouse databases. In addition, cavity hitlists generated using CavBase may also be viewed.
- Most Relibase+ searches can be started from the following buttons on the Relibase+ menubar:
  - **Text Search**: Allows searches on entry code, text (HEADER, TEXT, COMPND and SOURCE records), author name, ligand compound name and ligand entry code searches
  - **Sequence Search**: Allows you to search on amino acid sequence
  - **Smiles Search:** Allows you to retrieve ligand SMILES strings
  - **Sketcher**: The drawing area allows searches on 2D ligand substructure searching, 3D ligand substructure searching and nonbonded (protein-ligand or protein-protein) interaction searching
  - **Hitlists**: The hitlist manager allows you to view and combine results from previous Relibase+

searches
  - **Stored Results**: View the results of saved searches.
  - **Help**: This links through to the Relibase+ Help pages which includes both the Relibase+ and Reliscript User Guides, and the Relibase+ Technical Documentation
- Some Relibase+ searches can only be started from Protein Information pages or from Ligand Information pages, these are:
  - Ligand similarity searching.
  - Similar chain searching.
  - Similar binding site searching (and superposition).
  - WaterBase access.
  - CavBase access.

1.4 Relibase+ Text-Based Searches

1.4.1 Performing an Entry Code Search

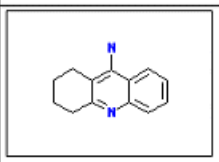
- Click on the **Text Search** button in the Relibase+ menubar.
- Type 1acj into the *Search String* box. PDB entry searches are exact match searches which will match on the given string only, i.e. a search on *pdb1et* will not retrieve *pdb1etr*.



- Hit the **Submit** button under to the *Search String* box to start the search.
- The results are presented as a single *Protein Information* page; the protein will be displayed in the AstexViewer at the top of the page.
- The protein structure can be analysed in more detail by launching Hermes. To do this, click on the **Show in Hermes** button in the *Hermes Controller* part of the Protein Information page:

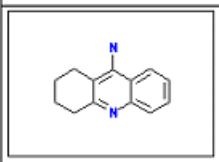
Whether or not the viewer is updated when navigating through Relibase+ can be controlled using the **Automatic Visualiser Updates** check box.

- A summary of the textual, bibliographic and crystallographic information for the entry is given in the body of the page:

Header	HYDROLASE(CARBOXYLIC ESTERASE)
Title	Unavailable
Compound	ACETYLCHOLINESTERASE (E.C.3.1.1.7) COMPLEXED WITH TACRINE
Reference	PROC.NAT.ACAD.SCI.USA **90**, (1993), 9031
Author(s)	J.L.SUSSMAN M.HAREL I.SILMAN
Source	(TORPEDO CALIFORNICA)
Method	X-RAY DIFFRACTION
Crystal	Cell: a = 113.700, b = 113.700, c = 138.100, alpha = 90.000, beta = 90.000, gamma = 120.000, Spacegroup = P3121
Resolution	2.800
RFactor	19.5%
Deposition Date	18.8.1993
Ligands	 THA in PDB-Entry 1acj **Water-Mediated Protein-Ligand Contact Paths: 2**
Protein Chains	pdb1acj-- Show sequence Submit Search for Similar Chains 100.0 Minimum Sequence Identity 100.0 Maximum Sequence Identity Save in Hitlist **Overwrite Existing Hitlist** Show Ligands
WaterBase	Water Information
CavBase	Cavity Information

- Detailed information on the water structure in the entry can be accessed by clicking on the **Water Information** button.
- Detailed information about cavities present in the protein are found by selecting the **Cavity Information** button. It is also possible to perform cavity similarity searches by clicking on this button.
- Additionally, at the top of every protein entry page there are several buttons:
  - **View PDB Header** - launches a browser window with the complete header of the original PDB file.
  - **Save PDB File -** allows the entire PDB file to be downloaded and saved.
  - **PDB Website** - links to the current entry on the PDB website (*http://www.rcsb.org/pdb/*).
  - **Bookmark** - allows you to save the page so you can return to it later.
- Open Hermes by selecting the **Visualise** button at the top of the page and inspect the structure found. Familiarise yourself with the various visualiser options. More information about Hermes can be found elsewhere.
- The GUI of Relibase+ is designed to allow easy navigation through the available data. This is realised *via* hyperlinks, which can be presented, for example, as ligand diagrams. Go from the protein page to the ligand page for the ligand bound to this protein by clicking on the ligand 2D

diagram for 1acj. This links you through to the *Ligand Information* page where the ligand will be displayed in AstexViewer. Hermes is updated automatically if the **Automatic Visualiser Updates** check box is activated.

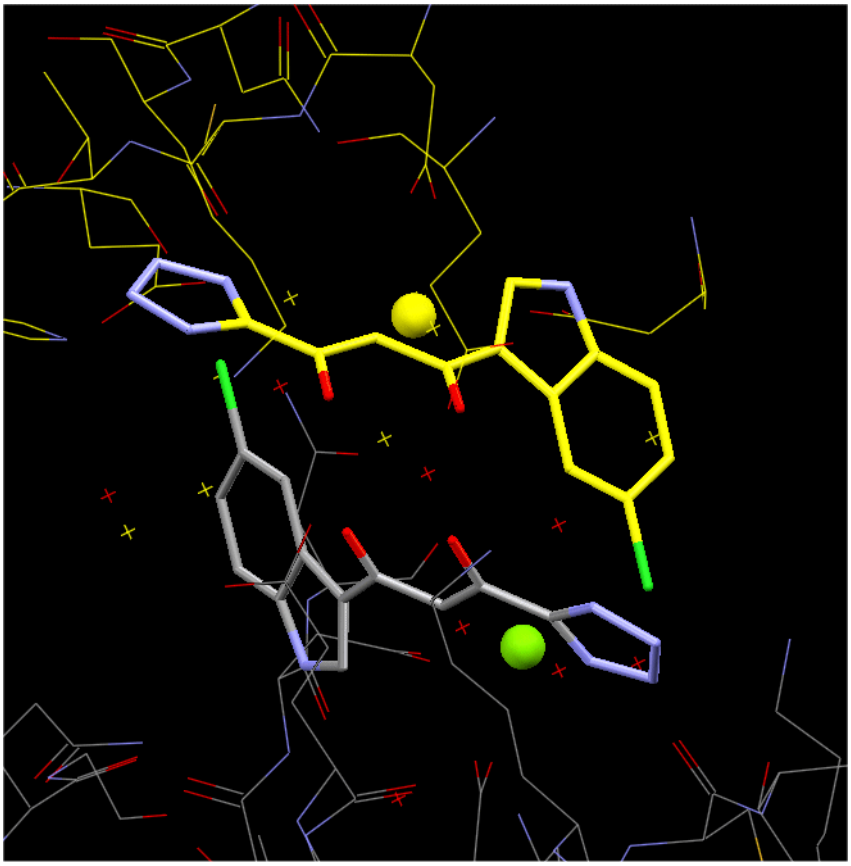
Header	HYDROLASE(CARBOXYLIC ESTERASE)
Title	Unavailable
Compound	ACETYLCHOLINESTERASE (E.C.3.1.1.7) COMPLEXED WITH TACRINE
Reference	PROC.NAT.ACAD.SCI.USA **90**, (1993), 9031
Author(s)	J.L.SUSSMAN M.HAREL I.SILMAN
Source	(TORPEDO CALIFORNICA)
Method	X-RAY DIFFRACTION
Crystal	Cell: a = 113.700, b = 113.700, c = 138.100, alpha = 90.000, beta = 90.000, gamma = 120.000, Spacegroup = P3121
Resolution	2.800
RFactor	19.5%
Deposition Date	18.8.1993
Ligands	 THA in PDB-Entry 1acj **Water-Mediated Protein-Ligand Contacts:** **Ligand** / **Water** / **Protein** / Visualise N15 (THA) / 634 / O (TRP 84) OG (SER 122) / show N15 (THA) / 643 / OD2 (ASP 72) O (SER 81) / show
Protein Chains	pdb1acj-- Show sequence Submit Search for Similar Chains 100.0 Minimum Sequence Identity 100.0 Maximum Sequence Identity Save in Hitlist **Overwrite Existing Hitlist** Show Ligands
WaterBase	Water Information
CavBase	Cavity Information

- Use the *Contacts* functionality in Hermes to see if the ligand forms any indirect hydrogen bonds to the protein (i.e. *via* a water molecule), and measure a few H-bonding distances. Also, highlight all the short nonbonded contacts in the binding site.

	Protein	H-Bonds	Short Contacts
1	THA_999_PDB1A	☑	☑

- Repeat the PDB entry code search for PDB entry 1qs4. Hyperlink through to the ligand information page and inspect the structure using Hermes as before. This time you will notice that there are columns present for *Metals* and *Packing* in the *Protein Explorer* window in Hermes.

- Packing is switched on by default; look at the packing of the ligand in the binding site:



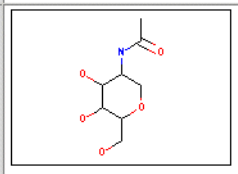
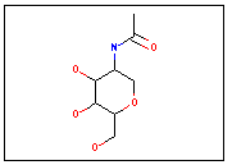
1.4.2 Text Searching

Text based searches are a convenient tool for retrieving sets of ligands or proteins. However, please keep in mind that due to inconsistencies in annotation, sequence or substructure searches are usually a better means of getting comprehensive information out of the database.

- In order to find a number of acetylcholinesterase structures click on the **Text Search** button in the Relibase+ menubar.
- Select **Keyword** from the *Search Type* pull-down menu then ensure the *Search Field* is **HEADER**, **TITLE**, **COMPND** and **SOURCE Records**.
- Type the required text string into the *Search String* box, i.e. *acetylcholinesterase*.
- Various options are available for all text-based searches:

- *Minimum MolWeight* and *Maximum MolWeight* boxes can be used to restrict the molecular weight (and size) of ligands that you wish to retrieve from your search. Leaving these boxes empty means that all ligands are considered.
- *Lowest Resolution* and *Highest Resolution* boxes can be used to restrict the experimental precision of the structures that you wish to consider. Leaving these boxes empty means that all structures are considered. If you only wish to consider structures with a resolution of 2.0Å or better then enter 2.0 into the *Highest Resolution* box. The resolution of NMR structures is set to -1.0Å, by default in Relibase, so entering 0.0 into the *Lowest Resolution* box will exclude all NMR structures from the search.
- *Structure Method Filters*: the **X-ray** and **NMR** check boxes can be used to filter searches based on the structure determination method.
- *Use Hitlist* allows you to restrict a search to a previously saved hitlist, which can be selected from the pull-down menu next to *Use Hitlist*. *Use Hitlists* will only be available if you have carried out a search and saved the hitlist.
- *Save in Hitlist* allows you to save the results of a search in a hitlist; type the required hitlist name into the *Save in Hitlist* box before you start the search. Previously saved searches can be overwritten by typing the appropriate search name into the Save in Hitlist box and activating the *Overwrite Existing Hitlist* check box, otherwise you will be prompted to give the search an alternative name.
- *Use Databases* allows you to select which database or combination of databases is searched. The databases must first have been loaded when the Relibase+ server was started (see Section 3 of the Inhouse Data Processing manual). The default setting is to search *All databases*.

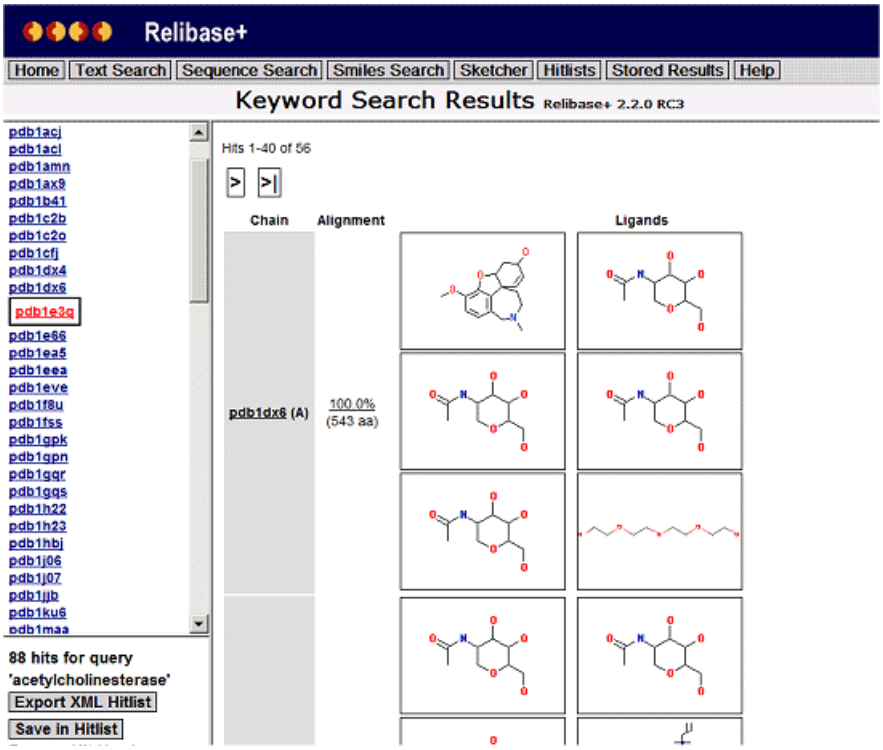
- Hit the **Submit** button to start the search. The results are presented as a browsable list of Relibase+ entries.
- The keyword searched for is highlighted in red.
- Browse thought the hits, select the **Browse Hits** hyperlink. From the resulting list search and locate the structure of acetylcholinesterase from Torpedo Californica (an electric ray), e.g. PDB entry 1e3q.

Header	SERINE HYDROLASE
Title	TORPEDO CALIFORNICA **ACETYLCHOLINESTERASE** COMPLEXED WITH BW284C51
Compound	MOL_ID: 1 MOLECULE: **ACETYLCHOLINESTERASE** CHAIN: A EC: 3.1.1.7
Reference	ACTA CRYSTALLOGR.,SECT.D **58**, (2002), 1765
Author(s)	C.E.FELDER M.HAREL I.SILMAN J.L.SUSSMAN
Source	MOL_ID: 1 ORGANISM_SCIENTIFIC: TORPEDO CALIFORNICA ORGANISM_COMMON: PACIFIC ELECTRIC RAY ORGAN: ELECTRIC ORGAN TISSUE: ELECTROPLAQUE
Method	X-RAY DIFFRACTION
Crystal	Cell: a = 114.000, b = 114.000, c = 138.200, alpha = 90.000, beta = 90.000, gamma = 120.000, Spacegroup = P3121
Resolution	2.850
RFactor	Unavailable
Deposition Date	20.6.2000
Ligands	 NAG in PDB-Entry 1e3q  NAG in PDB-Entry 1e3q **Water-Mediated Protein-Ligand Contact Paths: 1**

1.4.3 Similar Chain Searching

Acetylcholinesterase consists of only one polypeptide chain. To find all proteins with an amino acid chain that is identical to the one you are looking at (i.e. the Torpedo Californica acetylcholinesterase):

- Scroll down the 1E3Q protein information page to the *Protein Chains* part.
- The different protein chain sequences can be displayed by clicking on the appropriate chain hyperlink e.g. **pdb1a3q-A**.
- Select the **Submit** button next to *Search for Similar chains.* Tip: searches for similar or identical protein chains can be invoked from any protein page. The *Minimum Sequence Identity* and *Maximum Sequence Identity* boxes can be used to specify the required sequence identity as a percentage with respect to the reference chain (default is 100%).
- If you wish to display the ligand 2D chemical diagrams in the resulting list of chains, select the **Show Ligands** check box.
- The results are displayed as a table of chains, ranked according to their sequence identity relative to the reference chain:

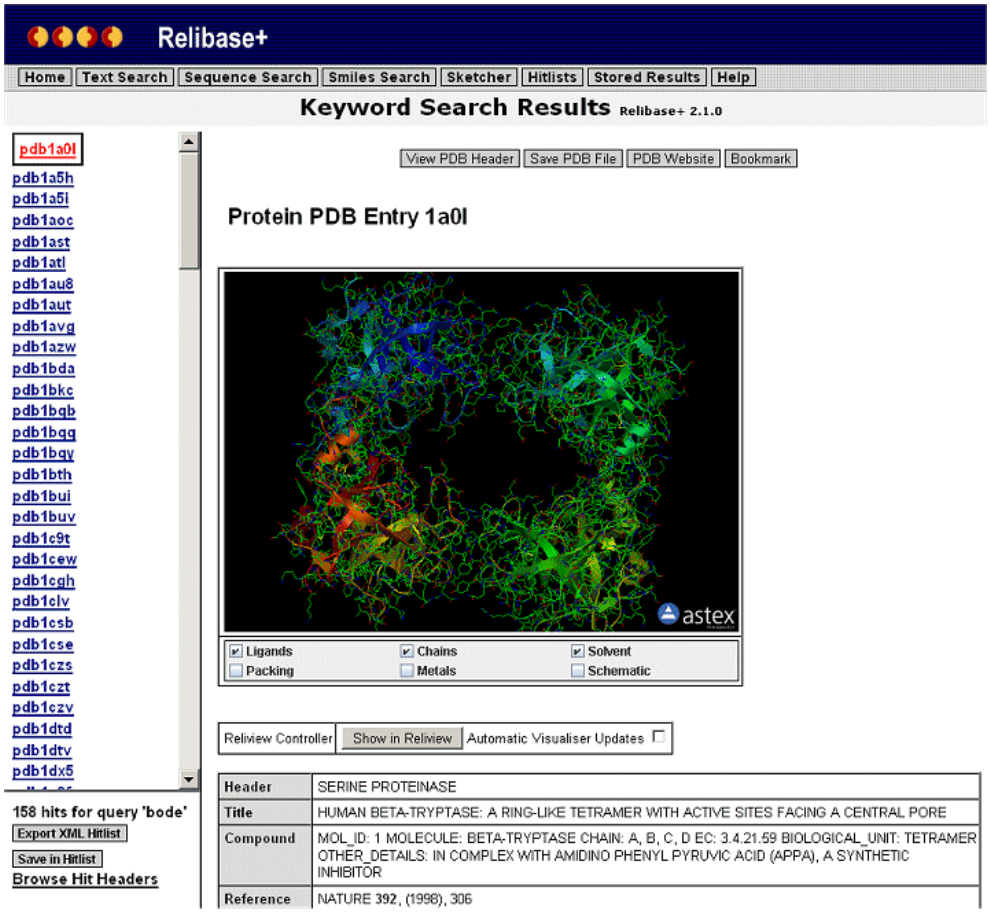


- The links from the % sequence identity in this table will show the alignment of two complete chains.
- Conserved residues are coloured blue, residues that are similar are coloured red, and residues that are completely different are coloured black.
- Check a few of the hits to see if the protein was isolated from the same type of electric ray.

1.5 Hitlist Manager

Hitlists allow storage of query results on the Relibase+ server. There are two types of hitlists, protein and ligand, depending on the type of input query. For example, author searches result in protein type (PDB entry code) hitlists, whereas searches for ligand names result in ligand type hitlists.
To find all structures in the PDB where *Bode* is one of the authors and then store the results in a hitlist:

- Click on the **Text Search** button in the Relibase+ menubar.
- Change the *Search Type* option from **PDB Entry Code** to **Keyword**.
- Type *Bode* into the *Search String* box.
- Ensure the *Search Field* pull-down menu reads **Author Name**.
- Enter a hitlist name, e.g. *Bode*, into the *Save in Hitlist* entry box.
- Hit the **Submit** button at the bottom of the page to start the search. The results are presented as a browsable list of Relibase+ entries:

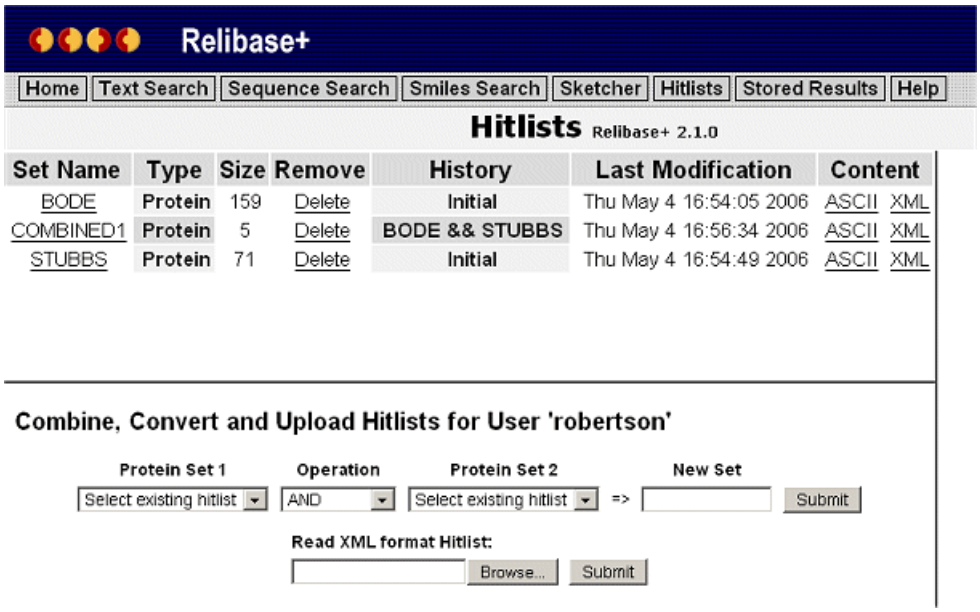


To see if any of the retrieved entries were also done by author *Stubbs,* carry out an author search, as just described, on *Stubbs* and save these results in a hitlist as well, e.g. *Stubbs*.

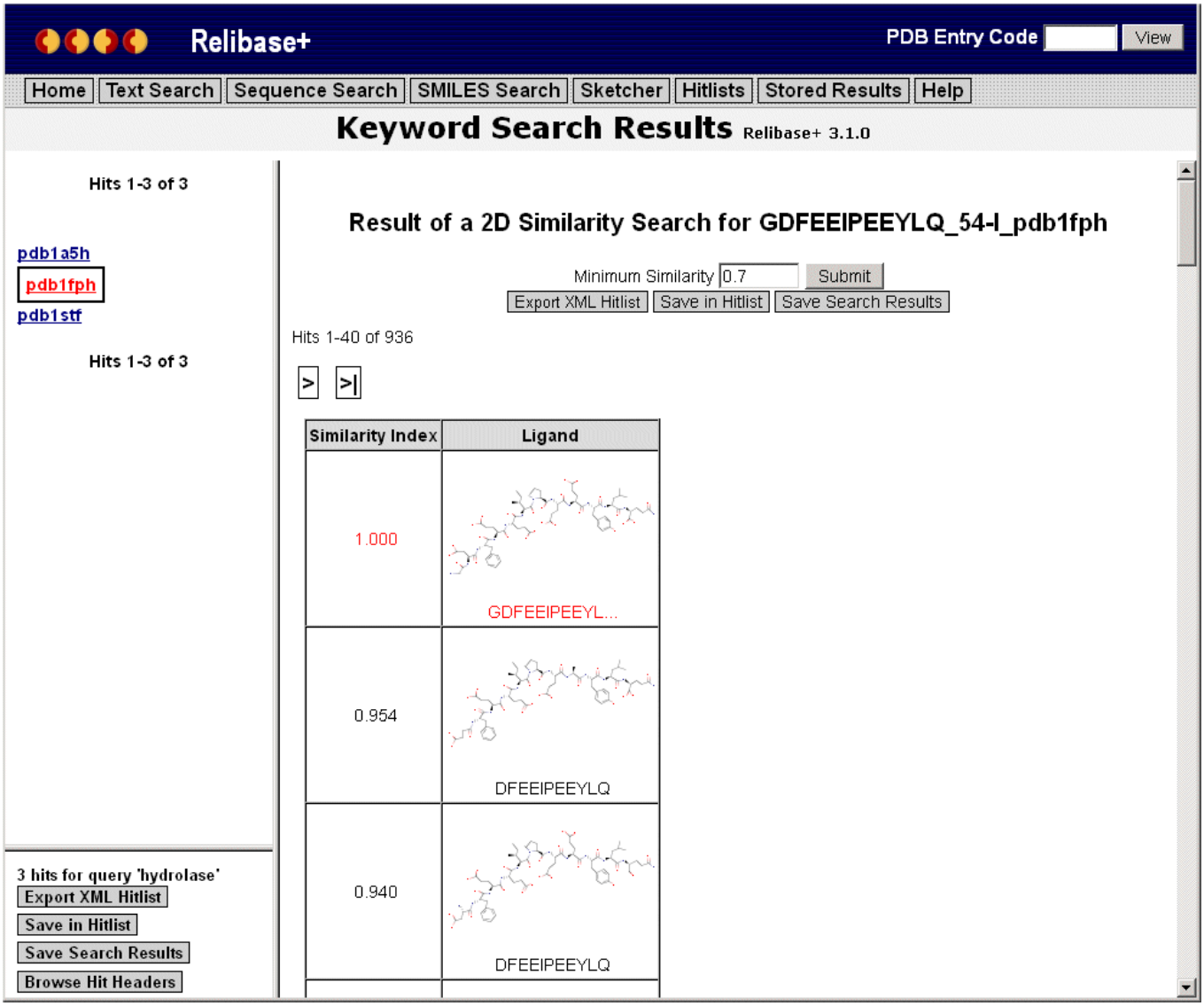
The Relibase+ hitlist manager can be invoked from the top level menubar, and this allows easy combination of hitlists using boolean operators. Thus, the intersection of two hitlists can be easily generated.

- Click on the **Hitlists** button on the Relibase+ menubar. This loads three frames into the browser, below the menubar. The top-left frame lists the hitlists you have stored according to *Set Name*, *Type* (Ligand or Protein) and *Size* (number of entries in hitlist). The *Remove* column allows hitlists to be deleted. *History* and *Last modification* date and time are also provided. If *Initial* is displayed in the *History* column, then no changes have been made to the saved hitlist. If, for example, a hitlist has been converted from a hitlist of ligand entries to a list of PDB entry codes then HITLISTNAME L-P would be given as the history.
- You should see the hitlists you have generated from the above searches displayed. Click on the name of the hitlist, e.g. *BODE* and the protein-ligand entry codes are displayed in the right-hand frame. Selected proteins can be added to a different hitlist or removed from the current, or another, hitlist by selecting the appropriate buttons.
- Combine the 2 hitlists you have generated, *BODE* and *STUBBS,* by selecting *BODE* from

Protein Set 1 and *STUBBS* from *Protein Set2*. Select **AND** from the list of boolean operators and type in a name for the new hitlist, e.g. *COMBINED1*. Finally, hit **Submit** to generate the new hitlist:



- Searches can also be restricted to a hitlist generated in a previous step. See how many structures published by both *Bode* and *Stubbs* are hydrolase structures.
- Click on the **Text Search** button in the Relibase+ menubar. Select **Keyword** as the *Search String* and type *hydrolase* into the *Keyword Search* box, ensuring the *Search Field* menu is set to **HEADER, TITLE, COMPND and SOURCE Records**. Select your combined author hitlist name e.g. COMBINED1 from the *Use Hitlist* entry box then hit **Submit**. Inspect the hits, for example, 1fph.
- Relibase+ also allows searching for similar ligands using topological fingerprints. These searches can be invoked from any *Ligand Information* page. Find out if the ligand in the thrombin structure 1fph was ever used as a ligand in another structure in the PDB and all the ligands which are most closely related to this ligand.
- From the GDF *Ligand Information* page of 1fph click on the **Similar Ligands Search** button at the top of the page. All ligands in the Relibase+ database are compared to the reference ligand (highlighted in red).
- The results are loaded into the browser as a list of ligands and are ranked in decreasing order of similarity to the reference ligand. The similarity index given in the first column is a Tanimoto coefficient and the 2D diagrams in the second column are linked to the corresponding ligand pages. If the search results in only one hit then you will be linked directly to the corresponding *Ligand Information* page:



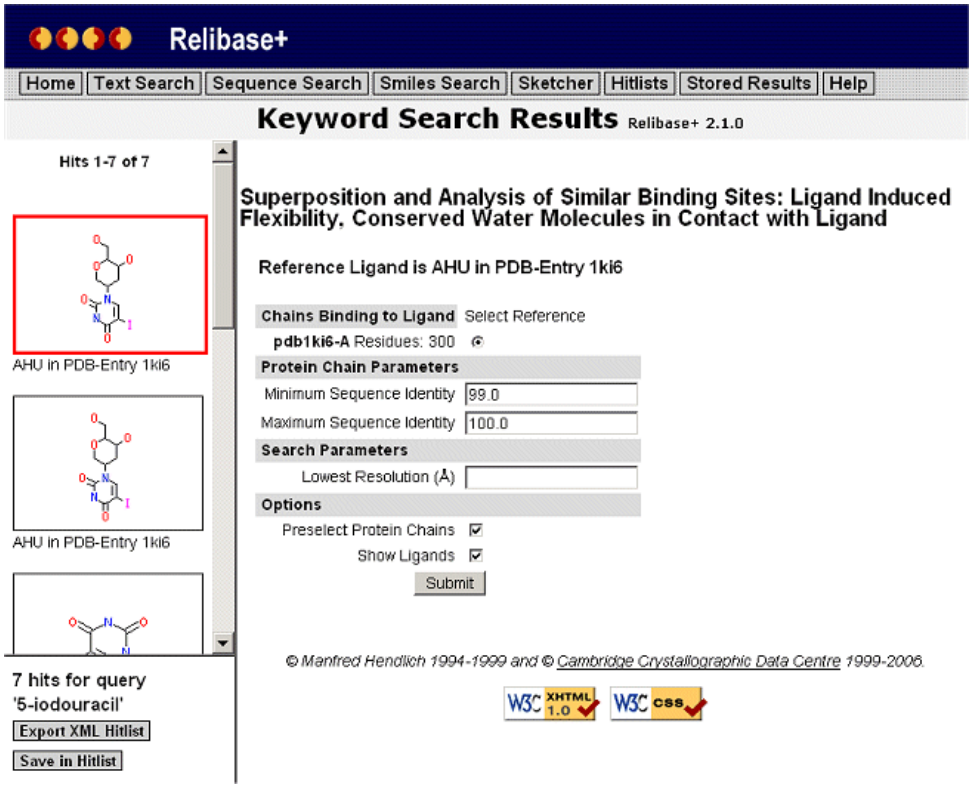
- The search results can be filtered on the basis of the Tanimoto coefficient (the default value is 0.7). Enter the required minimum similarity index (a value between 0 and 1) into the *Minimum Similarity* window and hit the **Submit** button.
- The hitlist of similar ligands can either output as an XML format file or saved on the Relibase+ server or using the **Export XML Hitlist** or **Save in Hitlist** buttons respectively.

1.6 Similar Binding Site Searches

Relibase+ allows superposition of similar ligand binding sites if the proteins share a significant level of sequence similarity. This two step approach can be used to easily analyse the common features and differences between those binding sites, such as protein flexibility, water conservation, ligand clashes etc.

- Find ligands containing a 5-iodouracil moiety using text-based searching.

- Load the ligand page for one of the two binding sites in the PDB entry 1ki6 which is a thymidine kinase complex. The starting point for this type of query is always a ligand page. There are eight buttons at the top of the *Ligand Information* page:
  - **Similar Ligands Search** - launches a search for ligands similar to that on the current ligand page.
  - **Similar Ligands in CSD** - launches a search for similar ligands in the Cambridge Structural Database, to that on the current ligand page.
    Note: this search is subject to the user holding a current CSD System licence.
  - **Similar Binding Sites Search** - launches a search for binding sites similar to that on the current ligand page.
  - **Save Mol2 File** - pops up a separate browser window with a Sybyl Mol2 text file of the ligand.
  - **Save SDFile** - opens a separate browser window with an SD text file of the ligand.
  - **Save Complex PDB File** - pops up a separate browser window with a PDB text file of the binding site.
  - **Save Complex Mol2 File** - opens a separate browser window with a Mol2 text file of the binding site.
  - **Bookmark** - allows you to save the page so you can return to it later
- In order to investigate similar binding sites click on the **Similar Binding Sites Search** button at the top of the *Ligand Information* page. This will launch a form which will allow you to find all binding sites which are at least 99% identical to the one you're looking at. In this example there is only one chain in close proximity to the binding site.
- Change the sequence identity limit in the *Minimum Sequence Identity* text box.
- If you want the ligand diagrams to be displayed in the resulting table of chains, make sure the **Show Ligands** check box is selected.
- If you wish all chains included in the 3D superposition to be preselected in the list of results ensure that the **Preselect Protein Chains** check box is switched on (this is the default); otherwise switch off this check box and then make your selection from the list of results.



- Start the search for similar chains by clicking on the **Submit** button. The results are displayed as a table of chains, ranked according to their sequence identity relative to the reference chain.
- Superimpose all the resulting binding sites. By default, all aligned residues in the chains are used initially for superposition, then 40% are removed for a refined superposition. Optionally, if you only want to use the binding site residues from that chain, make sure that the **For superposition Use Binding Site Residues Only** check box is selected. Further options are available. Click on the **Submit** button to superimpose the selected chains.
- The binding site superposition is displayed in the AstexViewer window. The superposition results can also be saved in mol2 format using the **Download Superimposed Structures** hyperlink beneath the 3D display.
- Textual results of the binding site superposition are tabulated. Detailed information on a particular binding site can be accessed by following the relevant hyperlink in the *Protein Chain* section.
- Superposition results can be saved by going to the *Save Superposition Results* part of the page (at the bottom). The results can be returned to via the **Stored Results** menu button. Alternatively, the results can be saved to a hitlist in the *Save in Hitlist* part of the page (also at the bottom).
- If the **Automatic Visualiser Updates** tick box is not enabled, view the results of the superposition by clicking on the **Show in Hermes** button.
- Use Hermes to have a detailed look at all components of the various superimposed complexes. Hermes can be used to:

- Switch (parts of) each binding site on and off.
- Colour the (various parts of) each binding site separately.
- View H-bond or short range interactions.

- There are two distinct binding modes depending on the type of ligand. Find out what characterises these two binding modes. Look at the positions of the water molecules to help find these.
- Investigate the results shown in the analysis of superimposed chains summary table. Investigate how flexible the main and side chains are. Are there any conserved water molecules in the binding site? Consider changing the reference by selecting the **Change Reference Chain** button and select the **Recalculate Table** button to create a new table.

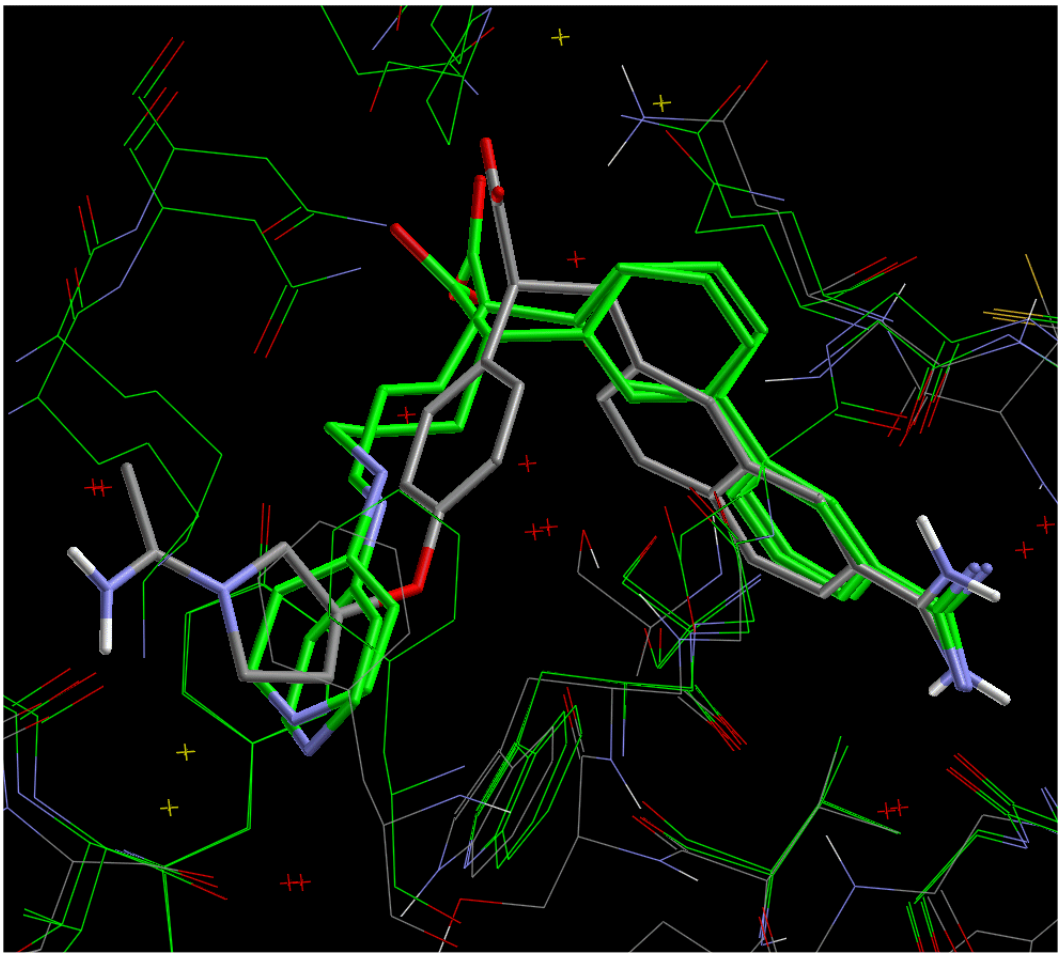
1.7 Investigating Crystal Packing Effects

superposition of binding sites can be carried out including crystal packing effects, thus allowing investigation of the influence of crystal packing on the ligand binding mode, within a series of related structures.

The aim of this example is to search for crystal packing effects in factor Xa binding sites and find out which structures are influenced by crystal packing effects:

- Click on the **Text Search** button in the Relibase+ menubar.
- Type 1fax into the *Search String* box, ensuring the *Search Type* menu is set to *PDB Entry-Code Search*, then hit the **Submit** button.
- Go to the Ligand Information page for DX9 in the retrieved PDB entry for 1fax and click on the **Similar Binding Sites Search** button as in the previous example.
- In the resulting page accept the default values and start the search for similar chains by clicking on the **Submit** button.
- In the options which are available for chain selection click on With **Ligands Only,** which will exclude entries with no ligands from your current selection.
- Since you want to investigate crystal packing for these similar binding sites ensure that the **Get Crystallographic Environment** check box is selected; **Packing** tick boxes will then be present in the Hermes *Protein Explorer.*
- Click on the **Submit** button to superimpose the selected chains.
- Use Hermes to have a detailed look at all components of the various superimposed complexes. As the **Get Crystallographic Environment** check box was selected there is an additional column of *Packing* check boxes which allow you to switch the crystal packing of the protein-ligand binding sites on and off. If there are no check boxes this indicates crystal packing isn't available.
- If you select 1xka you will see that packing residues (Lysine side chains) bind in the S4 specificity pocket which is filled up by parts of the ligand in all other superimposed structures in

this example. This indicates competition between the packing residues and the ligand in terms of interactions with the S4 specificity pocket. Docking calculations carried out using GOLD indicate that it is possible for the ligand in 1xka to bind in a more extended conformation filling up the S4 pocket. Competition for the S4 pocket is shown below:



- Investigating crystal packing in this way provides a very quick and easy method of assessing which crystal structures are likely to exhibit packing effects simply by scrolling though the 3D visualiser toolbox and looking at the available **Packing** check boxes.
- This can be of importance when looking at protein-ligand complexes and assessing the binding sites in structure-based drug design. If crystal packing is present then you need to assess if the ligand conformation is the physiologically relevant one, and is not distorted as a result of crystal packing effects. This also has an application in protein-ligand docking when using, e.g., GOLD and the incorrect ligand conformation is predicted when crystal packing effects are not taken into consideration. Including crystal packing allows GOLD to predict the correct ligand conformation.
- Since the reference structure, 1fax, does not contain any water molecules it cannot be used to carry out the analysis for conserved water positions. By using the **Change Reference Chain** button you can select a structure containing water molecules thus enabling the analysis to be

performed.

2 Tutorial 2: Substructure Searching in Relibase+ (SMILES and 2D/3D)

2.1 Introduction

A key feature of Relibase+ are substructure-based 2D and 3D searches. Searches for ligand substructures can be carried out using the SMILES search form. More complex searches, in addition to simple searches, can be set up in the Relibase+ 2D/3D query sketcher.

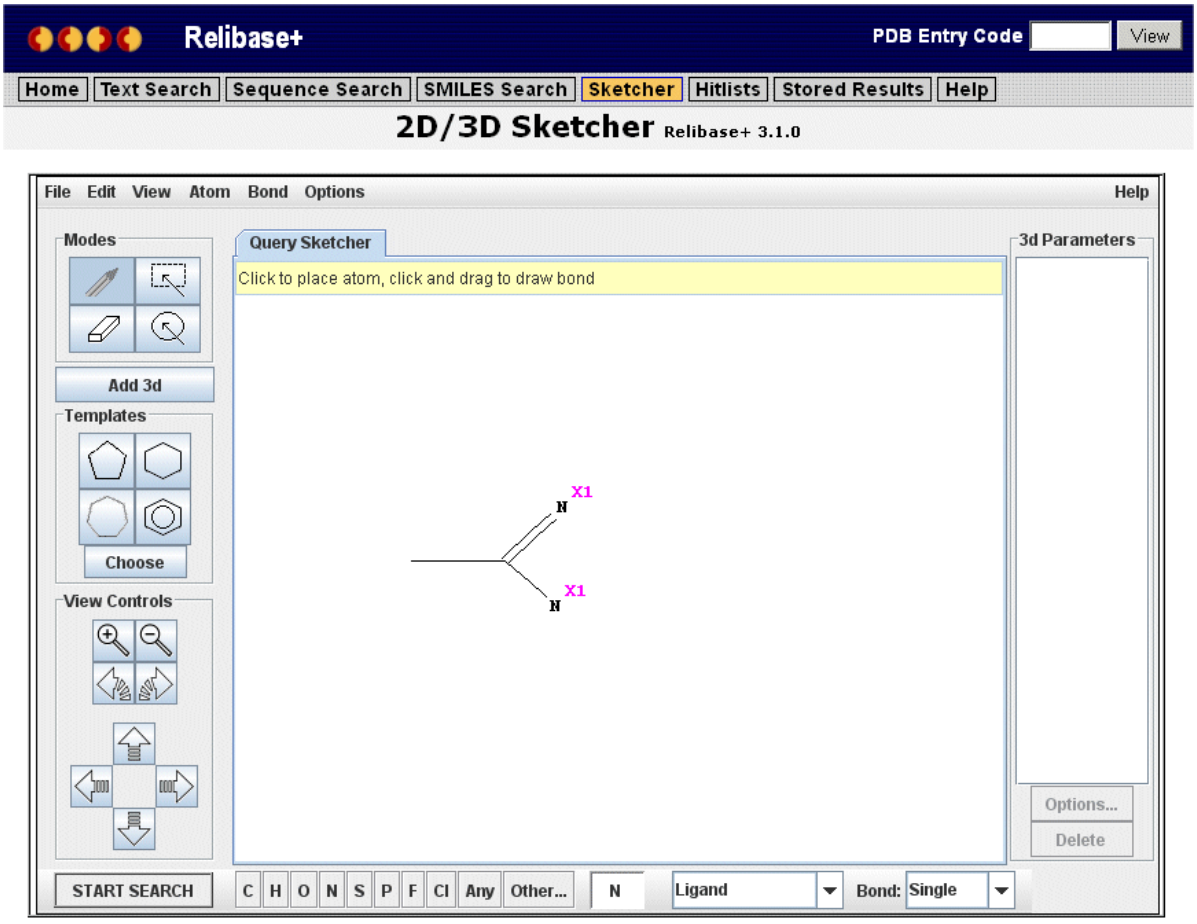
2.2 2D Ligand Substructure Searching (SMILES or 2D/3D)

2.2.1 Search for Ligands Comprising Different Substructures

The following search shows how to find ligands comprising two or more different substructures. A typical example could be the following task: Thrombin has very distinctive specificity pockets. The S1 pocket is well suited to accommodate basic moieties, such as amidino groups, whereas the S3/S4 aryl binding site preferably binds to large aromatic residues, such as naphthyl groups. Try to find all ligands containing both an *amidino* and a *naphthyl* group using both SMILES strings and the 2D/3D sketcher:

- Click on the **Smiles Search** button in the Relibase+ menubar.
- Type the required amidino Smiles string into the *Enter SMILES Code* box: *[Cc]C(=[ND])[ND]* and save the results of this search in hitlist: *amidino.*

- Click **Submit** to run the search.
- Repeat the search for naphthyl and type the required Smiles string for naphthyl into the *Enter SMILES Code* box: *c1cc2ccccc2cc1* and save the results of this search in hitlist: *naphthyl.*
- Click on the **Hitlists** button on the Relibase+ menubar and combine the two hitlists you have generated, *AMIDINO* and *NAPHTHYL*. Generate a new hitlist using the *AND* operator by selecting *AMIDINO* from *Ligand Set 1* and *NAPHTHYL* from *Ligand Set 2*, e.g. *NAPAM.*
- Note down the numbers of hits in each of these hitlists, including the new one you have just created combining the two hitlists.
- Look at some of the hits to observe what type of hits are being retrieved.
- Click on the **Delete** link next to *NAPAM* in the column labelled *Remove* to remove the combined hitlist.
- Repeat the substructure searches using the 2D/3D sketcher.
- Click on the **Sketcher** button on the Relibase+ menubar to take you into the Relibase+ sketcher.
- Ensure that the molecule type is set to **Ligand** and draw the amidino group attached to a carbon atom, as shown below:



- To specify that the N atoms are only connected to one non-hydrogen atom, right click on the first N atom and select **Number of connections to non-hydrogen atoms**, **1** from the pull-down, then do similarly for the second N atom.
- Click on the **Search button**. This will take you to the **Start Search** window where a number of filters and other search options can be defined, including the search name.
- Select the **Hitlist Controls** tab and type *amidino* into the *Save search in hitlist named:* box.
- By default only 1000 hits will be returned for this search. To return all substructure matches select the **Hit Limits** tab and select the **Show all hits** radio button.
- Hit the **Start button** to initiate the search; the amidino hitlist saved from the SMILES string search will be overwritten by default. Relevant hits are displayed in the *Results* window during the search. A new browser window containing all the hits is launched when the search has completed.
- Return to the sketcher by selecting the **Query Sketcher** tab. Delete the amidino fragment by clicking on **Edit** and **Delete All**.
- Ensure the molecule type is set to **Ligand** and select the *Naphthalene* template by clicking on the Other button in the **Templates** part of the interface, **Ligand**, **Unsaturated Rings** and then **Naphthalene**. The group is loaded by clicking in the drawing area. Click **Search** and type *naphthyl* into the *Save search in hitlist named:* box in the *Start Search* window. Click on **Start** to run the search.
- Combine the hitlists generated as described earlier recreating *NAPAM* again and ensure that the same number of hits are retrieved for each hitlist are the same using the different substructure searching methods.

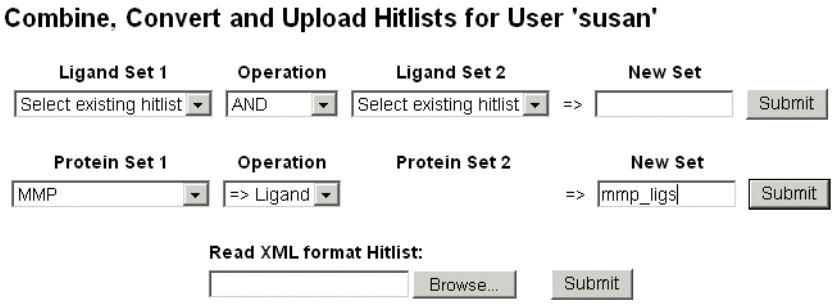
2.2.2 Combining Search Methods

This search is also about combining various search methods. The aim of this example is to try and find all organic ligands of MMP-3 (stromelysin-1), which is a matrix metallo-protease containing a zinc atom in its active site. To do this select all ligands bound to MMP-3 sharing 100% sequence identity with stromelysin-1: PDB-entry 1ums. This will demonstrate the ability of the hitlist manager to translate ligand type hitlists into protein type hitlists:

- Click on the **Text Search** button in the Relibase+ menubar.
- Type 1ums into the *Search String* box and ensure the *Search Type* menu is set to **PDB Entry-Code Search**. Hit the **Submit** button to start the search.
- Scroll down to the bottom of the Protein Information page retrieved for 1ums to *Search for Similar chains*. Save the results of the similar chain search in a hitlist: *mmp* and then select the **Submit** button next to *Search for Similar Chains* to start the search.
- The results are displayed as a table of chains, ranked according to their sequence identity relative to the reference chain.
- Now click on the **Smiles Search** button in the Relibase+ menubar.
- Type *C* into the *Enter SMILES Code* box, which will allow us to find all organic ligands and save

the results of this search in hitlist: *organic*. A molecular weight restriction could also be applied to remove small organic ligands, such as acetate. The same search could easily be carried out using the 2D/3D sketcher as described before and the results saved in a hitlist. Hit the **Submit** button to start the search.

- Click on the **Hitlists** button on the Relibase+ menubar. We want to be able to combine the two hitlists that have been created, however, it's not possible to combine a ligand type hitlist with a protein type hitlist so one of the hitlists must be converted into another hitlist type. In this example the task was to find ligands so we will the convert the protein type hitlist *MMP* to a ligand type hitlist*: mmp_ligs*.
- Select the hitlist you wish to convert using the pull-down menu below *Protein Set 1*.
- Select => *Ligand* from the pull-down menu of logical operators.
- Enter the name of the new ligand set in the text box below *New Set: mmp_ligs*. Hit the **Submit** button adjacent to create the new hitlist. Now you have list of all ligands contained in MMP-3.



- Combine the two hitlists: *MMP_LIGS* and *ORGANIC* using the *AND* operator to create a ligand hitlist: *orgmmpligs* which correspond to the task required. Inspect the list to check that the structures expected have been found so far.

2.3 3D Interaction Searches

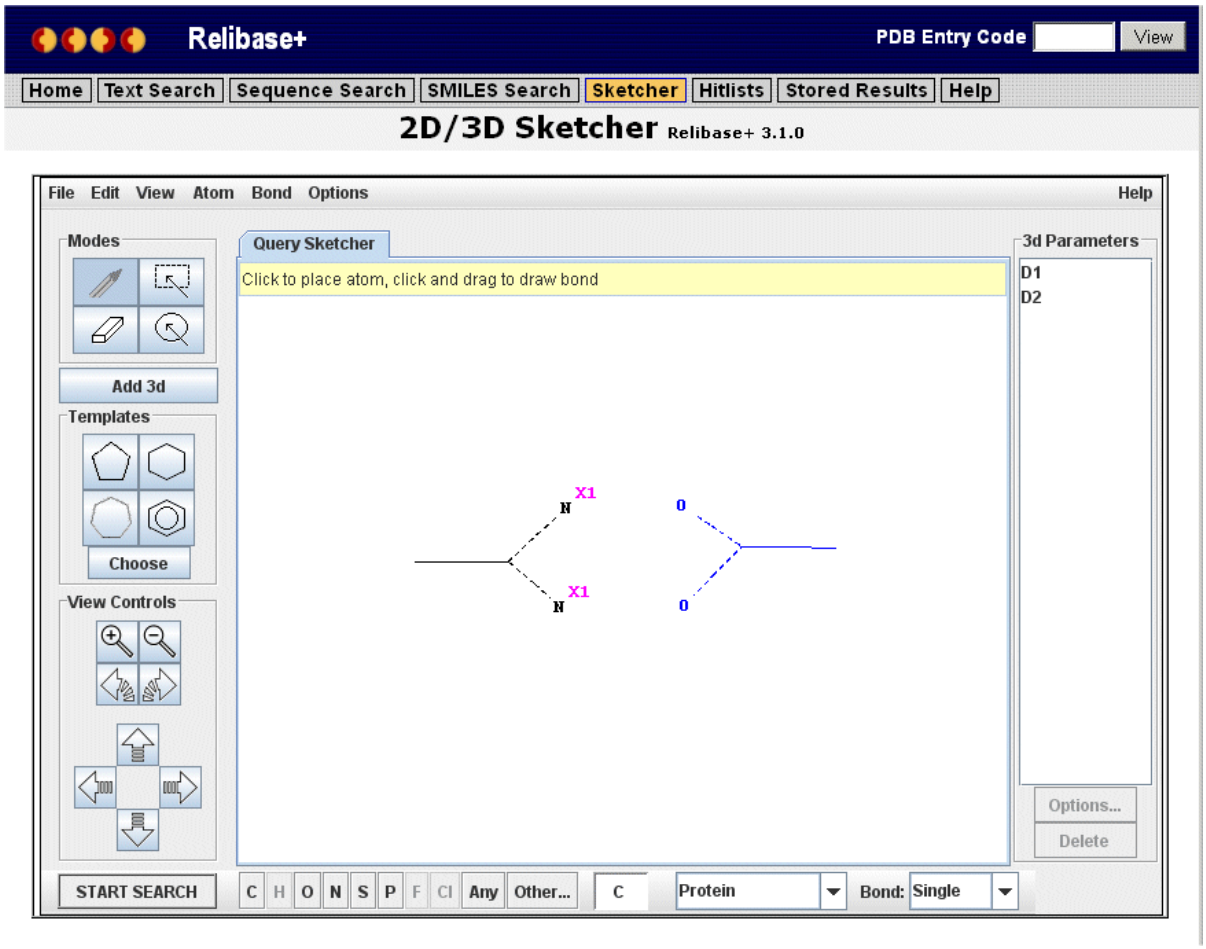
The following examples are designed to explore the abilities of the sketcher for setting up 3D searches, e.g. for particular protein-ligand interaction patterns. The sketcher supports three types of molecules: **protein**, **ligand**, and **water**, for specifying the respective types of fragments. Combinations of these types are also available, e.g. **Protein or Ligand**, **Protein or Water**, **Ligand or Water** in addition to an **Any** molecule type. Independent fragments must be correlated by applying constraints, i.e. distance, angle or torsion angle constraints.

2.3.1 Protein-Ligand Interaction Searching

Use the *AMIDINO* hitlist generated in the previous example and search for the ligands which form a salt bridge with a carboxylate group of a protein. Use distance constraints to find bidentate salt

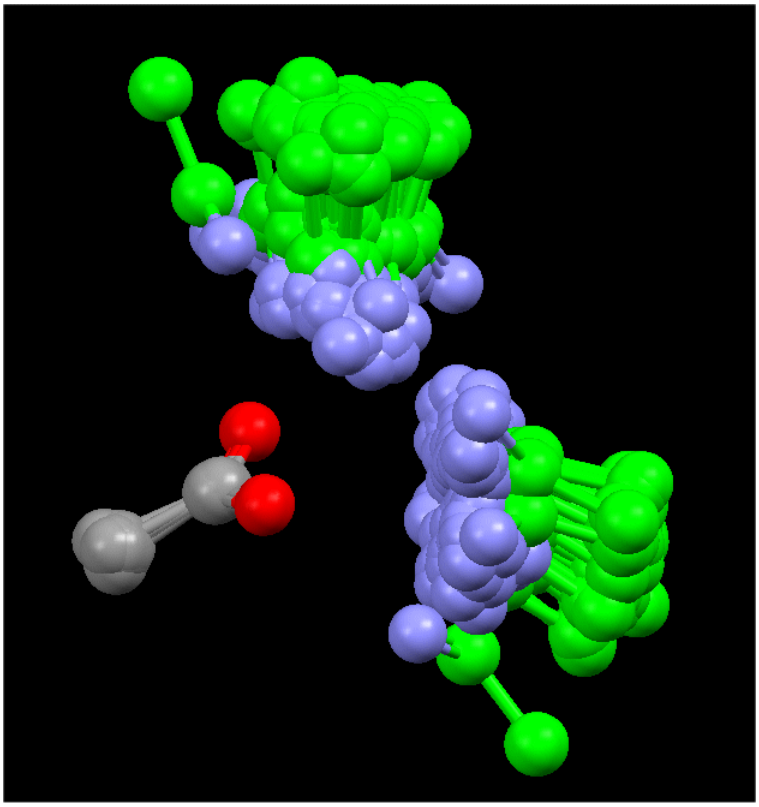
bridges:

- Click on the **Sketcher** button on the Relibase+ menubar to take you into the Relibase+ sketcher.
- Draw the amidino group attached to a carbon atom, as in the previous example, but this time use *Any* bond type for the carbon to nitrogen bonds and ensure that the molecule type is set to **Ligand.**
- Change the molecule type to **Protein** and draw the carboxylate group with *Any* bond type for the carbon to oxygen bonds.
- Set up distance constraints between the N and O atoms by clicking on the **Add 3d** button to the left of the drawing area. Select the atoms involved in the constraint, i.e. *N* and *O*, hit **Define** next to *Distance* in the *Add 3d* pop-up window. Once you have defined the first pair of atoms, do similarly for the remaining N and O atom.

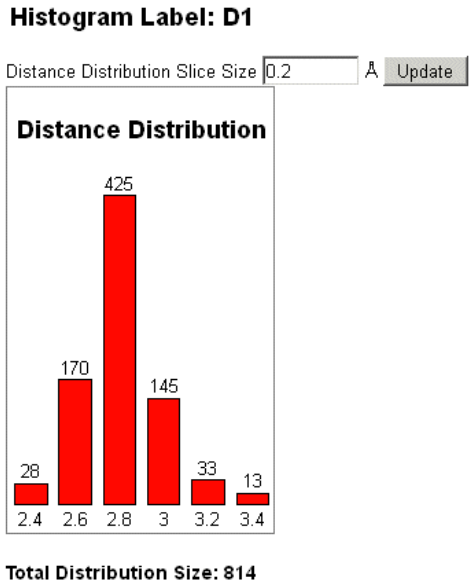


- Now define the angle between the planes of the two fragments. To do this first select the N-C-N atoms and hit the **Define** next button next to **Plane:** in the *Edit 3D Parameters* window. In the same way define the plane for the O-C-O moiety. To define the angle between these planes, select **Plane1** and **Plane2** in the *Edit 3D Parameters* window, then hit **Define** next to *Angle.*
- Click on **Done** to close the window.

- Hit **Search** to start the search. In the *Start search* window, go into the **Hitlist Controls** tab and type *AMIDINO* into the *Restrict search to hitlist named:* entry box and then *BIDENT* into the *Save search in hitlist named:* entry box; this will save all hits corresponding to bidentate salt bridges.
- Select the **Start** button to start the search. Look at some of the hits in Hermes to ensure that you have retrieved the expected results (click on the **Show in Hermes** button if *Automatic Visualiser Updates* is not activated): the atoms included in the parameter definition are indicated in the visualiser using a van der Waals surface.
- The results of the search (matching groups) can be superimposed onto a set of selected atoms. Relibase+ will use all atoms *selected* prior to submission of the query for superposition. Investigate how the amidino groups bind to the carboxylate group, look at the orientation of binding and use carboxylate as a reference for superposition; see which contact distance is the most common:
- Return to the query, ensure the sketcher is in *Select* mode and select 3 atoms of the protein carboxylate group by pressing *Shift* and left-clicking on the atoms. Alternatively, depress the left mouse button and drag a rectangle to encompass the 3 atoms of the carboxylate group. Selected atoms are shown in red in the drawing area. Click **Search**.
- In the *Start search* window, select the **Superimpose hits on selected atoms** check box to generate an overlay of the hits. Selection of this check box generates a pull-down menu from which you can choose to **Display matching atoms only**, **Display matching chains** or **Display entire binding site**. Keep the default of **Display matching atoms only** for this example.
- Hit **Start** to run the search.
- When the search is complete, the results are displayed in the main Relibase+ window. Click on the **View in Hermes** button to view the results in Hermes.



- Select the **Histogram(s)** link to find which contact distance is the most common. When you have defined any geometrical parameters in your query, histograms for these parameters are generated automatically and can be viewed by clicking on the **Histogram(s) link** in the bottom left frame of the Relibase+ ligand browser:



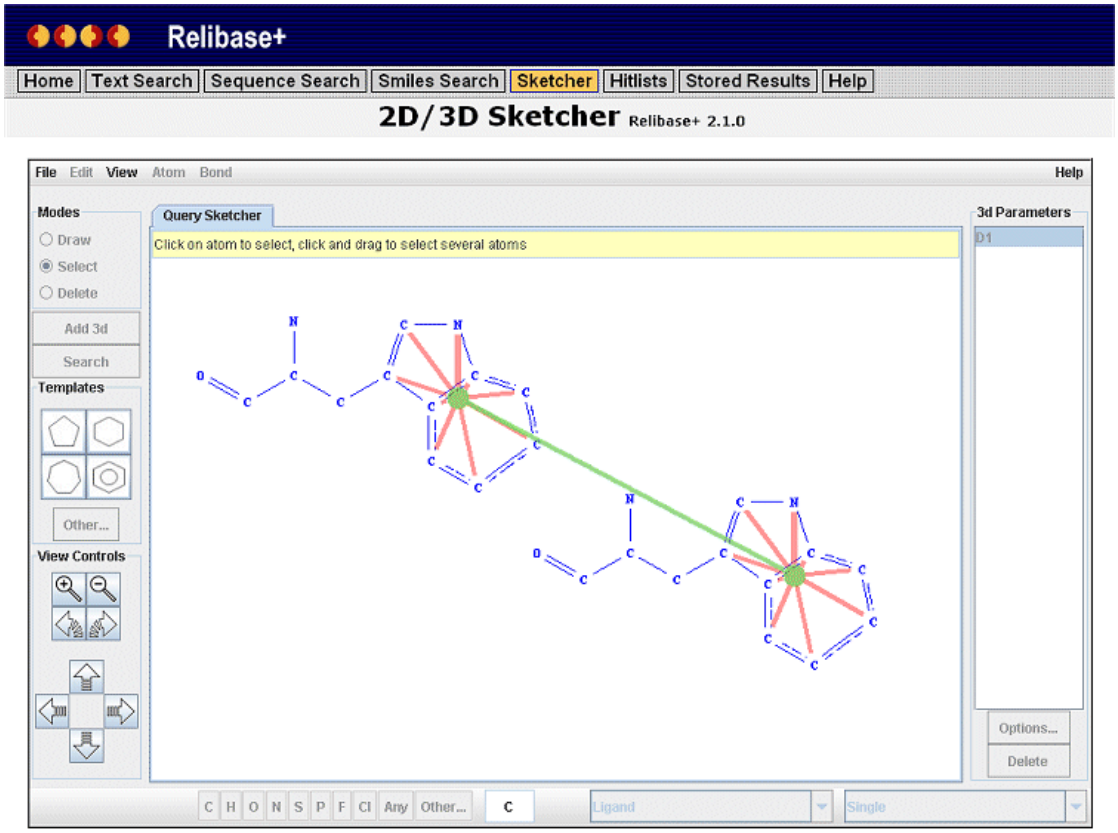
- We can also inspect the preferred angle between the planes of the two moieties.

- The histogram(s) can be exported in a comma-separated value (csv) file for analysis in third party software, and hyperlinks to structures from specific histogram bins are available.

2.3.2 Protein-Protein Interaction Searching

The Relibase+ sketcher also allows you to set up protein-protein interaction searches without specifying any ligand substructure. Centroids of selected atoms can be defined as a means to specify more complex geometrical situations. Find proteins containing two indole groups of tryptophan residues in close contact. Investigate one of the resulting hits while the search is still running. Describe the Trp-Trp interaction geometry of one of the hits:

- Click on the **Sketcher** button on the Relibase+ menubar.
- Click on the **Other** button in the *Templates* part of the Relibase+ interface, to the left of the drawing area, then select **Protein** and **Tryptophan** from the resultant pull-down menus. Load the template by clicking to the left of the drawing area.
- Define the centroid of the indole ring by clicking on the **Add 3d** button and selecting all the atoms that make up the ring (9 in total). This can be done more easily by depressing the left mouse button and dragging a rectangle which incorporates all nine ring atoms. Once all have been selected, click **Define** next to *Centroid:* in the *Add 3d* pop-up window. The text *CENT1* will be given in the *Defined objects* section of the *Add 3d* window. Click **Done** once you have finished.
- Load a second tryptophan template in the same manner, ensuring it is loaded to the right of the drawing area. Define the centroid of the indole ring in the same way as outlined before, but keep the **Add 3d** window open.
- Set up a distance constraint: left-click on *CENT1* and *CENT2*. Once both *CENT1* and *CENT2* are selected, click on the **Define** button next to *Distance.*



- Before closing the *Add 3d* window, select the distance *D1* and click on the **Options** button. This launches a dialogue box which will allow you to alter the limits on the distance constraint. Change the upper limit to *2.0* then click **OK** and then **Done** to finish the definition.
- Click **Search**.
- In the *Start Search* window specify that the search is to be restricted to X-Ray Structures only, with a *Lowest resolution* of 2.0 Angstrom, then click **Start** to initiate the search.
- Click on any of the entries that appear in the **Results tab** to view hits generated for this search. Results will be presented in a second browser window while the search is still running.
- Once the search is complete look at a hit, such as PDB entry 1cv2 which is a very nice example of the occurrence of pi-pi stacking interactions.

3 Tutorial 3: An Introduction to the Cavity Information Module and Hermes

3.1Overview

3.1.1 Objectives

To find ligands that might bind to a particular protein binding site. The binding site of interest will be used as the query cavity in a CavBase search, looking only for hit cavities that contain ligands. If a hit cavity is sufficiently similar to the query, then the ligand occupying that hit cavity might also bind to the query cavity.

3.1.2 Steps Required

Set up a cavity hitlist so that we only search a subset of the database.

- Set up the search query, specifying which parts of the query cavity we want to find matches for.
- Specify search options and run the search.
- View a hit cavity together with the query cavity in the 3D viewer, using various viewer options to assess how well the two match.
- Experiment with some other viewer settings.

3.1.3 The Example

The *S. aureus* multi-drug-binding repressor protein QacR is known to bind several drugs and is relevant to the mechanism of multi-drug resistance in this organism (D. S. Murray, M. A. Schumacher and R. G. Brennan. Crystal Structures of QacR-Diamidine Complexes Reveal Additional Multidrug-binding Modes and a Novel Mechanism of Drug Charge Neutralization. *J. Biol. Chem.* **279**, 14365-14371, 2004).

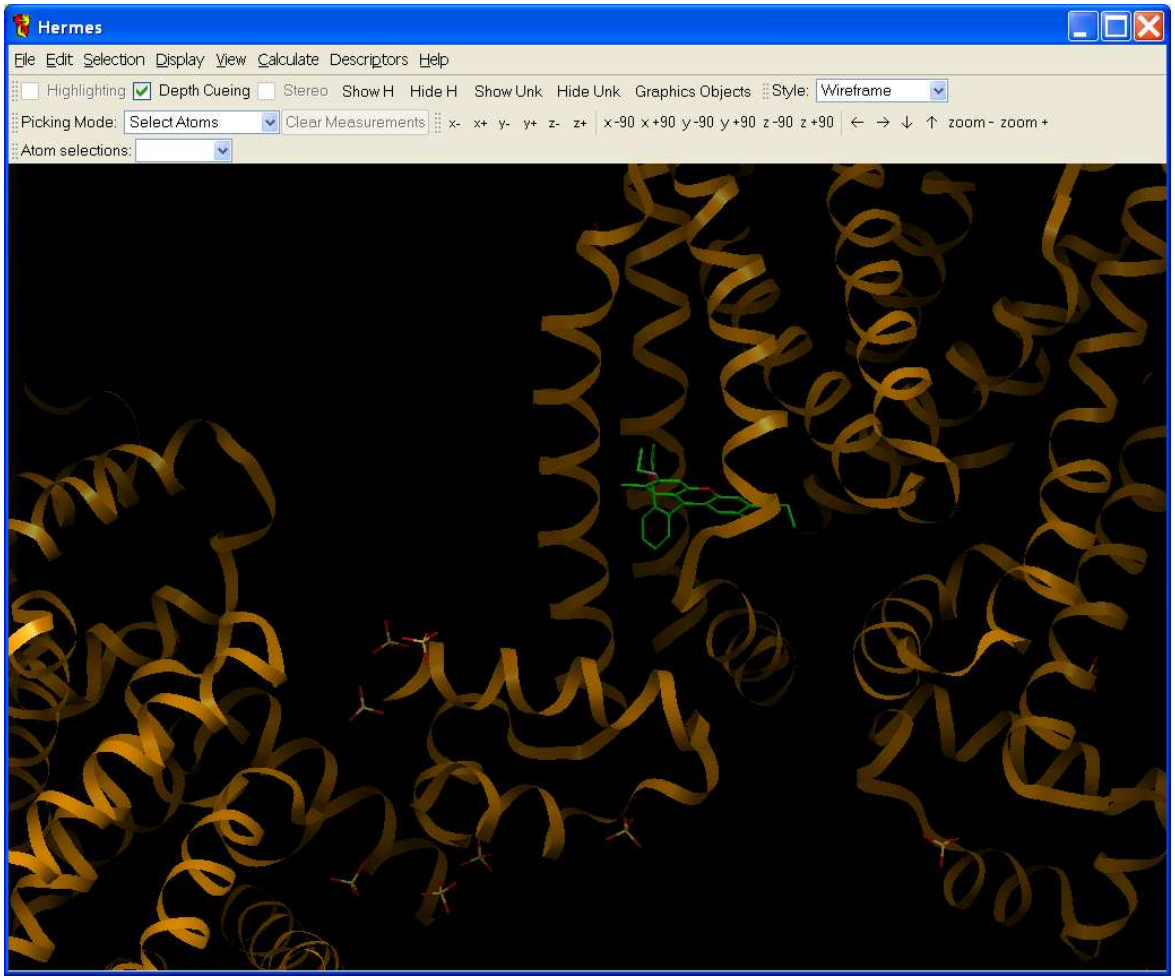
- In this tutorial, we will use CavBase to search for other ligands that might bind to this protein.

3.2 Setting Up a Hitlist

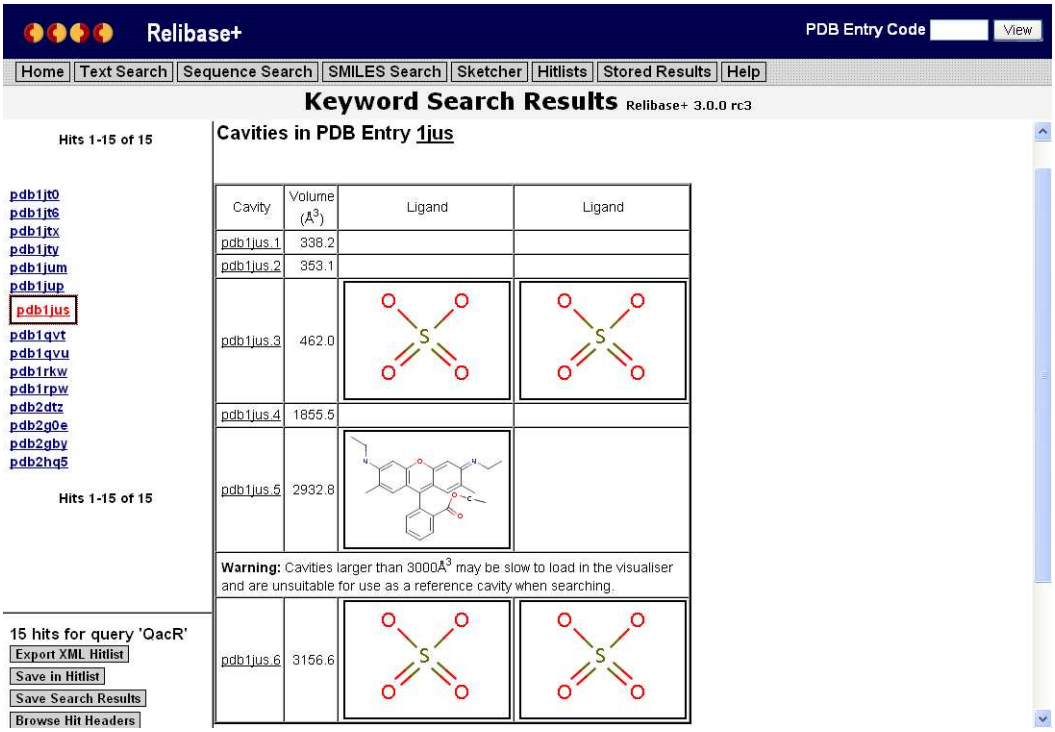
- Open Relibase+.
- When we do the cavity similarity search, we could, if we wished, search the whole database, but this would take a long time. To speed up the tutorial, we will therefore search on a subset that just contains the esterases. (There is no good reason for this except that, as we will see, there is at least one interesting hit in this subset!)
- To do this, we must first do a text search for *esterase* and save the hits in a protein hitlist.
- So, hit the **Text** button, select *Keyword Search* from the *Search Type* drop-down list, and type *esterase* into the *Search String* box. Enter a suitable name, e.g. *tutorial,* into the *Save in Hitlist* box. Hit the **Submit** button to run the search.
- When the search is finished, hit the top-level **Hitlists** button. There will be a *TUTORIAL* hitlist stored which we will use as a database subset for our cavity search.

3.3 Setting Up the Search Query

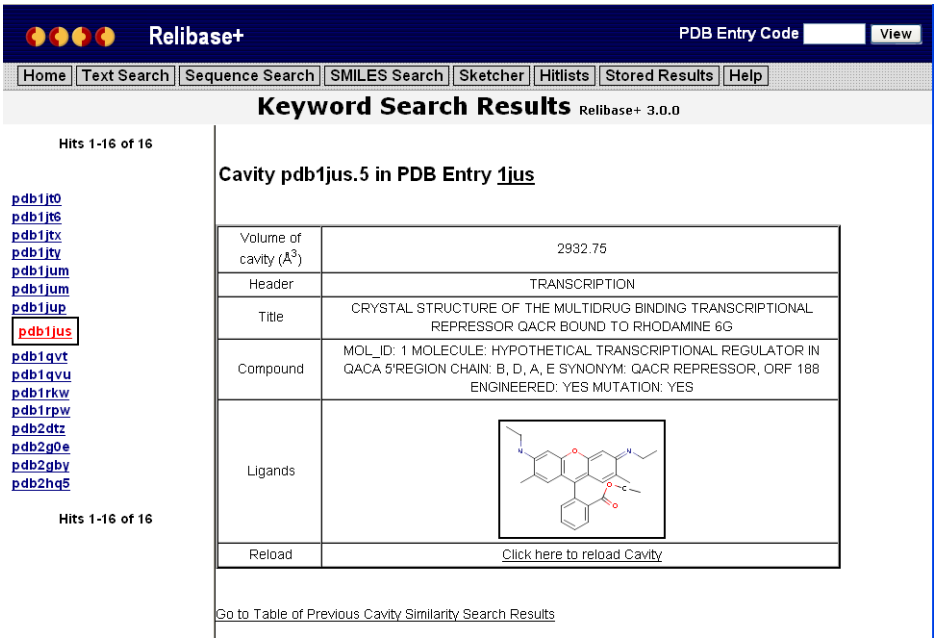
- Now we turn to the cavity similarity search. The first thing is to find the query protein cavity.
- Hit the top-level **Text** button, and do a keyword search for *QacR*.
- There are several relevant structures of which we will use just one. Click on *pdb1jus* in the list of hits to show the Protein Information page for this structure.
- Hit the **Show in Hermes** button located in the *Hermes Controls* Panel, this will launch the Relibase+ visualiser. As you see, this is a structure of the QacR protein with a bound ligand, rhodamine 6G:


Hermes
File Edit Selection Display View Calculate Descriptors Help
Highlighting Depth Cueing Stereo Show H Hide H Show Unk Hide Unk Graphics Objects Style: Wireframe
Picking Mode: Select Atoms Clear Measurements x- x+ y- y+ z- z+ x-90 x+90 y-90 y+90 z-90 z+90 zoom- zoom+
Atom selections:

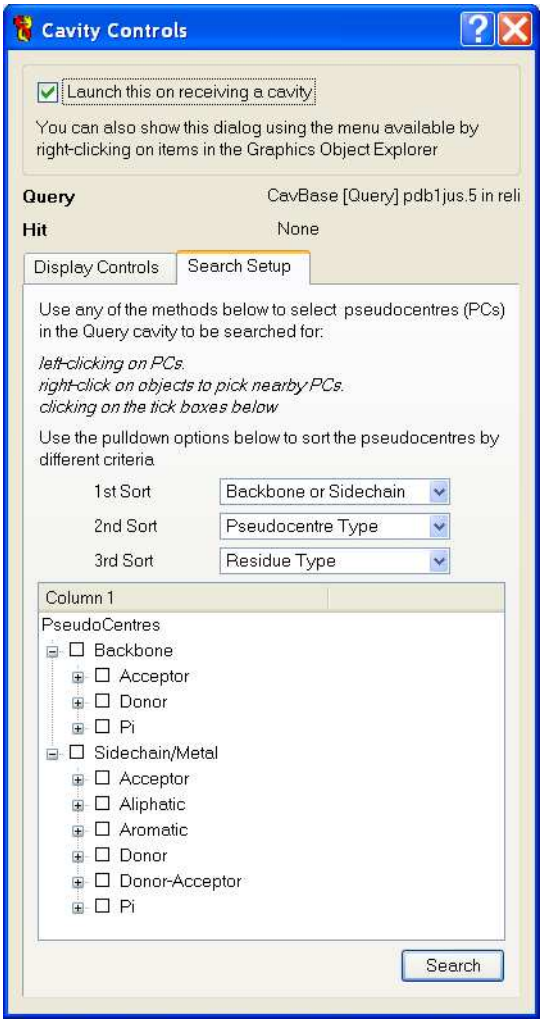
- Go to the very bottom of the webpage and click on the **Cavity Information** button.
- You now see a list of the cavities in this protein structure. We will use as a query the fifth cavity in the list, occupied by a rhodamine 6G ligand, viz. *CAV::pdb1jus.5*.



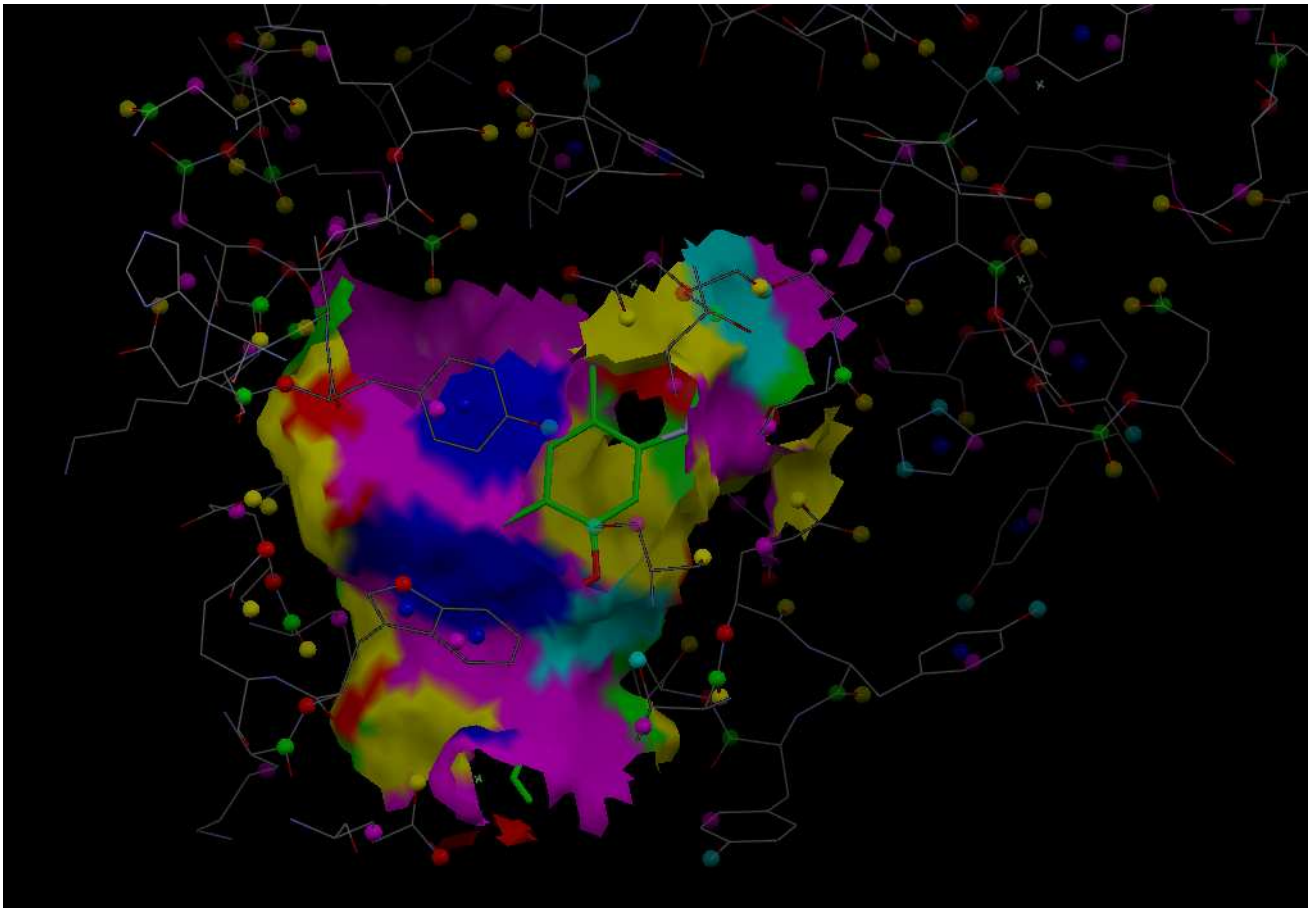
- Click on the *CAV:pdb1jus.5* link. This takes us to the cavity information page for this cavity:



- The cavity will be displayed in 3D within Hermes. The *Cavity Controls* window will also come up Go to the *Search Setup* pane of the viewer by hitting the right-hand tab within this window:



- To begin with, all pseudocentres in the query cavity are deselected and appear translucent. We need to choose which of them are to be included in the search query. A sensible strategy is to choose all the pseudocentres in the vicinity (say, within 5Å) of the ligand. To do this, click with the right-hand mouse button on any atom of the ligand, and pick **Select Pseudocentres within range of this ligand...** from the resulting pull-down menu. Type *5.0* in the resulting dialogue box and hit **OK**:
- The 3D display now shows the portion of the cavity that the search will try to match. The picked pseudocentres are depicted solid:

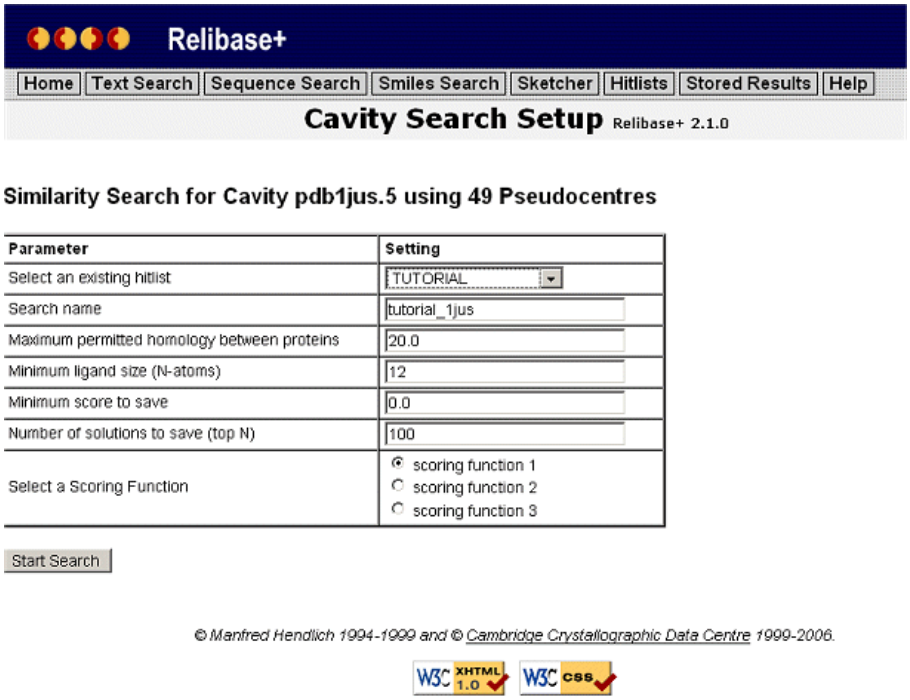


- Hit the **Search** button at the bottom right of the *Cavity Controls* window to complete the query definition. This should open a browser page that will enable us to set some other search options.

3.4 Selecting Search Options and Running the Search

- We do not want to search the whole database, so click on the down-arrow icon to the left of the *Select an existing hitlist* box and select from the resulting pull-down menu the hitlist that we created earlier, which you may have called *TUTORIAL*.
- Type in a search name, e.g. *tutorial_1jus*.
- Delete the *100.0* from the *Maximum permitted homology between proteins* box and type in *20.0* instead. By doing this, we ensure that any hits we find will have low sequence homology with QacR, i.e. we would not have been able to find them by more standard sequence-based methods.
- The point of this exercise is to find other ligands that might bind to QacR so, in this case, we only want to find cavities that are occupied by ligands (and ligands of a reasonable size). Delete the *0* from the *Minimum ligand size (N-atoms)* box and type in *12* instead. This means that we will not find any cavities that do not contain a ligand of at least 12 atoms.
- By default, the search will keep the 100 most similar cavities that it finds, irrespective of their score. Leave these settings as they are.

- We have a choice of scoring function. There has been little work so far to investigate which is best, so the choice between them is somewhat arbitrary. In this example, we will use scoring function 1.
- The dialogue should now look as follows:



- Start the search by hitting the **Start Search** button.
- After a few seconds, you will be shown a page that monitors the progress of the search. This page will be updated every 15 seconds. The search will take a few minutes to run.
- Once the search has run, you will be presented with a list of the hits that have been found. By default, they are ranked by similarity score, so the cavities that match the query best will come first. Many of the hits are Acetylcholinesterase. In the second page of results should be the hit *pdb2jf0.9*. This has very low sequence homology with the query (18.6%). Seven of the pseudocentres that you included in the search query were matched to pseudocentres in the *pdb2jf0.9* cavity. *Note: This cavity search is for illustrative purposes only. The scores and number of pseudocentres for the best hits matched in this search are quite low and so cavities that are very similar to the cavity in QaCR have not been found.*

Search Results 1-15 of 100

Cavity	Score	Normalised Score	Matched Centres	RMS	Protein Homology	Header	Title
pdb2v1o.12	8.0	18.1	9	0.443	17.4	HYDROLASE	CRYSTAL STRUCTURE OF N-TERMINAL DOMAIN OF ACYL-COA THIOESTERASE 7
pdb2gyu.12	6.4	14.6	7	0.772	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH HI-6
pdb2jey.12	6.1	13.8	8	1.210	18.6	HYDROLASE	MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH HLO-7
pdb2gyw.12	5.8	13.3	7	1.055	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH OBIDOXIME
pdb2ha3.13	5.8	13.2	6	0.821	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MOUSE ACETYLCHOLINESTERASE COMPLEXED WITH CHOLINE
pdb2c0q.13	5.8	13.2	6	0.492	18.6	HYDROLASE	NON-AGED FORM OF MOUSE ACETYLCHOLINESTERASE INHIBITED BY TABUN
pdb2h9y.13	5.8	13.2	6	0.791	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MOUSE ACETYLCHOLINESTERASE COMPLEXED WITH M-(N,N,N-TRIMETHYLAMMONIO)TRIFLUOROACETOPHENONE
pdb2ha5.12	5.8	13.2	6	0.756	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MUTANT S203A OF ACETYLCHOLINESTERASE COMPLEXED WITH ACETYLTHIOCHOLINE
pdb2jez.8	5.7	13.0	7	1.006	18.6	HYDROLASE	MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH TABUN AND HLO-7
pdb1n5m.14	5.6	12.8	6	0.550	18.6	HYDROLASE	CRYSTAL STRUCTURE OF THE MOUSE ACETYLCHOLINESTERASE GALLAMINE COMPLEX
pdb2jey.13	5.6	12.7	6	0.619	18.6	HYDROLASE	MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH HLO-7
pdb2jf0.9	5.5	12.6	7	1.979	18.6	HYDROLASE	MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH TABUN AND ORTHO-7
pdb1j06.12	5.5	12.6	6	0.787	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MOUSE ACETYLCHOLINESTERASE IN THE APO FORM
pdb2jgf.12	5.4	12.3	6	0.650	18.6	HYDROLASE	CRYSTAL STRUCTURE OF MOUSE ACETYLCHOLINESTERASE INHIBITED BY NON-AGED FENAMIPHOS
pdb1n5r.12	5.4	12.2	6	0.554	18.6	HYDROLASE	CRYSTAL STRUCTURE OF THE MOUSE ACETYLCHOLINESTERASE PROPIDIUM COMPLEX

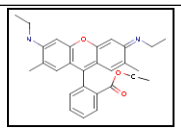
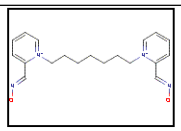
Browse the next 15 hits...

3.5 Viewing a Hit and Comparing It with the Query

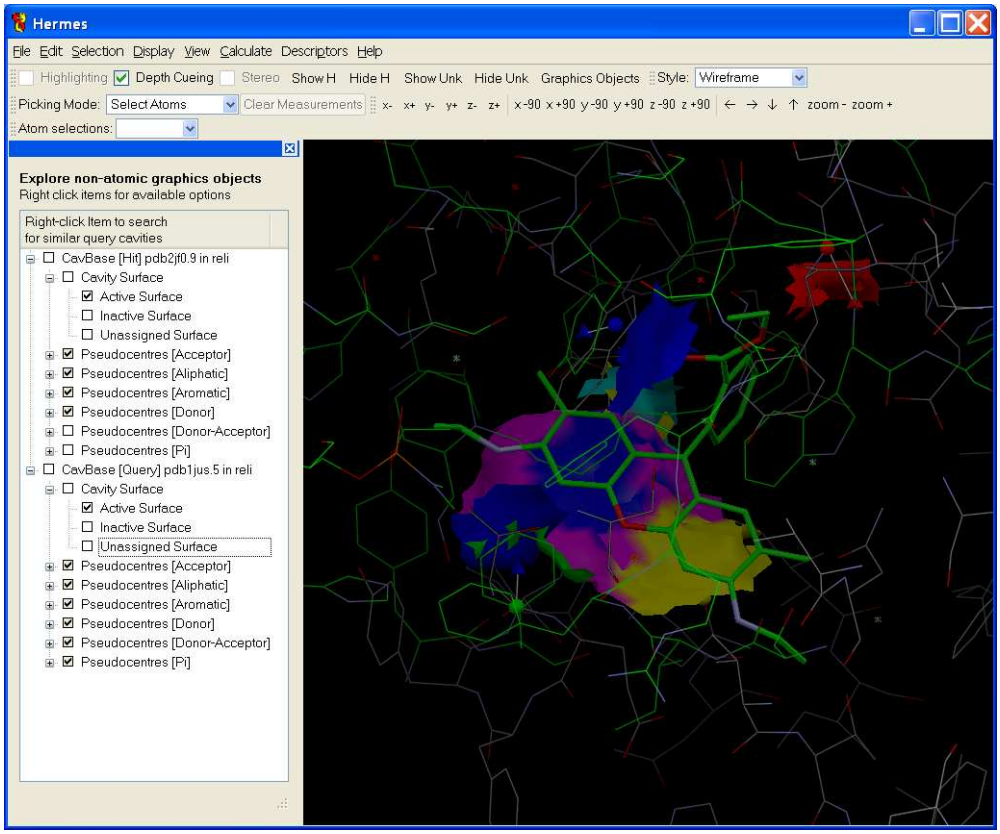
- To view this hit, click on the *pdb2jf0.9* link. This will display a browser page giving details of the hit. You will see that the hit is an acetylcholinesterase (AChE) structure.

Comparison of cavities pdb1jus.5 and pdb2jf0.9

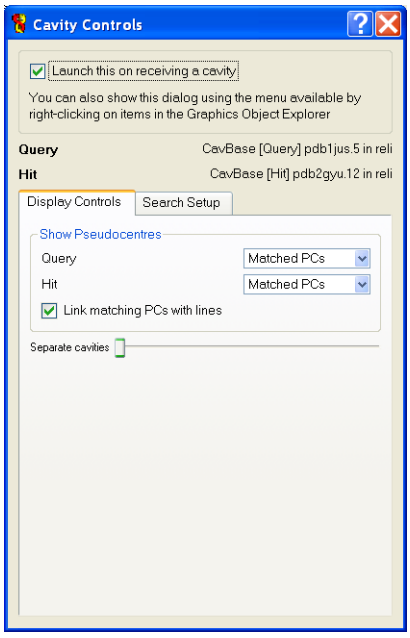
Pseudo-centres (match)	7
RMS (match)	1.98
Pseudo-centres (clique)	8
RMS (clique)	1.98
Score	5.55
Protein Homology	18.60
Visualiser Automatic Visualiser Updates ☐	Load in Hermes

	Query cavity pdb1jus.5		**Similar cavity pdb2jf0.9**
Volume ($Å^3$)	2932.75		1021.12
Pseudo-centres (total)	195		90
PDB	1jus		2jf0
Header	TRANSCRIPTION		HYDROLASE
Title	CRYSTAL STRUCTURE OF THE MULTIDRUG BINDING TRANSCRIPTIONAL REPRESSOR QACR BOUND TO RHODAMINE 6G		MUS MUSCULUS ACETYLCHOLINESTERASE IN COMPLEX WITH TABUN AND ORTHO-7
Compound	MOL_ID: 1 MOLECULE: HYPOTHETICAL TRANSCRIPTIONAL REGULATOR IN QACA 5'REGION CHAIN: B, D, A, E SYNONYM: QACR REPRESSOR, ORF 188 ENGINEERED: YES MUTATION: YES		MOL_ID: 1 MOLECULE: ACETYLCHOLINESTERASE CHAIN: A, B FRAGMENT: CATALYTIC DOMAIN, RESIDUES 32-574 SYNONYM: ACHE EC: 3.1.1.7, 3.1.1.1 ENGINEERED: YES
Ligands			
Cavity object	go to cavity information page		go to cavity information page

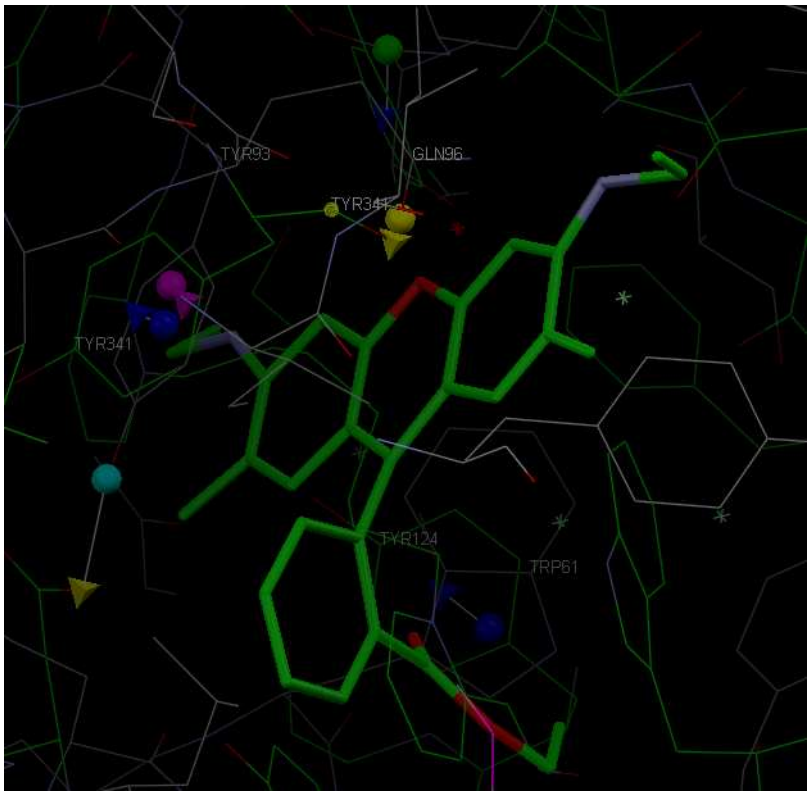
- Click on the *Load in Hermes* link. This will display both the query and the hit cavities superimposed and will also display the pseudocentres and surface patches that have been matched. Another window, the *Cavity Controls* window will also come up.
- You can hide the two window that appear to the left (The *Protein Explorer* and the *Contacts* windows) to increase the display area, by clicking at the top-right of each window. Click on the *Unassigned Surface* tick boxes in the window headed *Explore non-atomic graphics objects* (This is the *Graphics Object Explorer*) to hide surface patches unassigned. The display should be similar to that below:



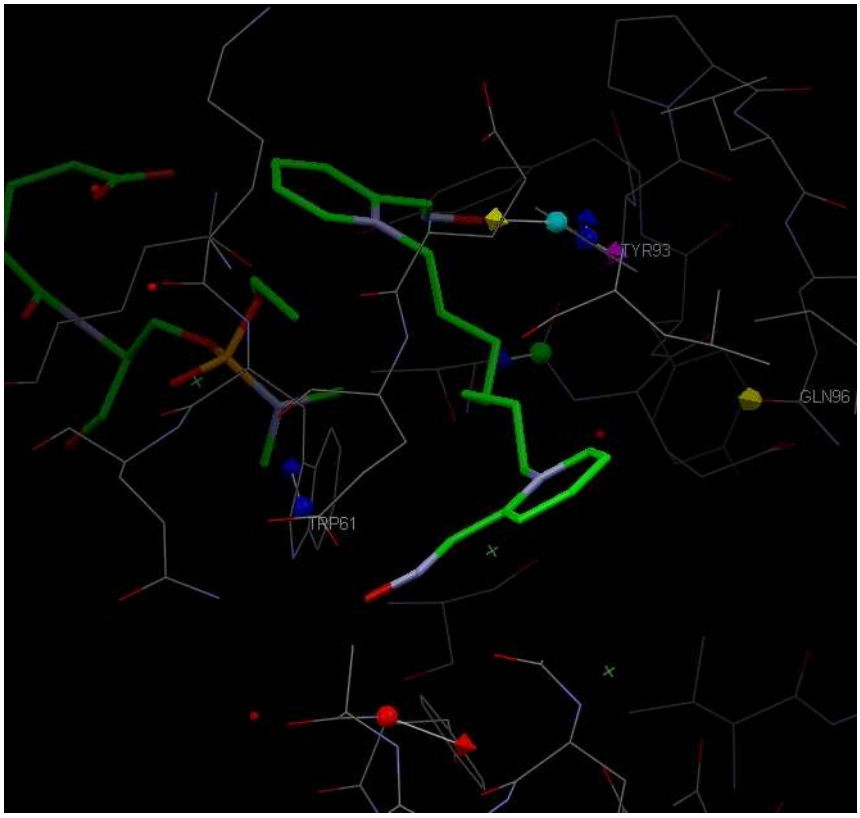
- Check that the pseudocentre display options in the *Display Controls* tabbed pane of the *Cavity Controls* window, are set to *Matched PCs* (which they should be, as this is the normal default setting):


Cavity Controls
Launch this on receiving a cavity
You can also show this dialog using the menu available by right-clicking on items in the Graphics Object Explorer
Query CavBase [Query] pdb1jus.5 in reli
Hit CavBase [Hit] pdb2gyu.12 in reli
Display Controls
Search Setup
Show Pseudocentres
Query Matched PCs
Hit Matched PCs
Link matching PCs with lines
Separate cavities

- Hide the surface patches for both cavities by clicking the *Active Surface* tick boxes in the *Graphics Object Explorer.*
- Only the matched pseudocentres will be displayed. The display should look as illustrated below, and shows that, amongst other things, Trp61 of the QacR cavity has been matched with Tyr 124 of AChE, and Tyr93of QacR with Tyr341 of AChE. (You can get the number of a residue by clicking with the right-hand mouse button on any atom of the residue and selecting **Labels** followed by **Label by Protein Residue**.)



- If you have labelled anything, remove the labels by clicking with the right mouse button anywhere in the display-area background and selecting **Labels** followed by **Do not label**.
- Now we will examine how well the ligand from the hit cavity might fit into the query cavity. You will need to activate the *Protein Explorer* window by selecting it from under **View** in the top level menu-bar of Hermes. Turn off the *Chain* tick box for the query and turn off the *Ligand* tick box for the Hit.
- The ligand from the hit cavity should now be displayed in the protein environment of the query cavity. The display should look something like this:



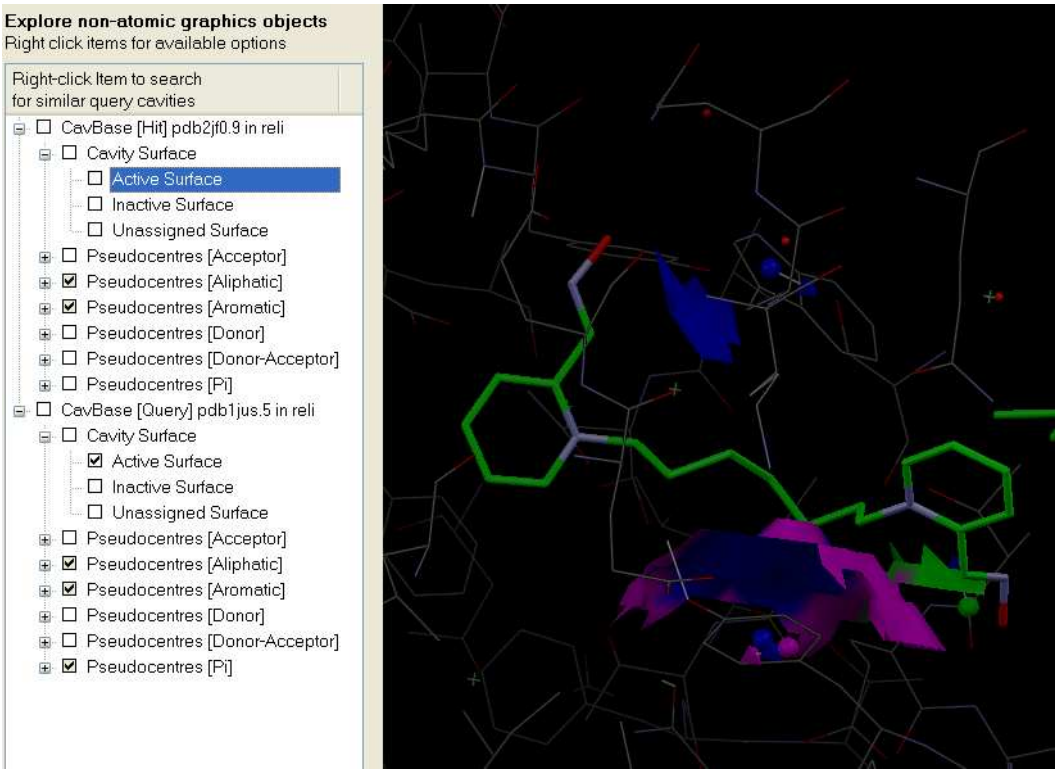
- The ligand (*ortho 7)* that resides in the hit cavity in acetylcholine esterase has an orientation in the query cavity that looks credible (the ligand has been made more prominant by displaying it in capped stick mode).
- To investigate the steric fit more closely, click with the right-hand mouse button on the *CavBase[Hit]* select **Styles** and then **Spacefill**.
- Whether or not this ligand would actually bind to QAcR is open to question, of course, but we are alerted to the possibility.

The tutorial can be finished at this point but, if you are interested, the next section will demonstrate a few more of the viewer options.

3.6 Experimenting with Other Visualiser Options

- The *Graphics Object Explorer* provides control of the display and the colour of pseudocentres and surfaces.
- First activate the active surface for the hit cavity by clicking the corresponding *Active Surface* tick box in the *Graphics Object Explorer.*
- Click with the right-hand mouse button on items in the *Graphics Object Explorer*, select **Colours** and select a colour from the pull-down menu to change the colour of a surface (this is not possible for an Active Surface), or a pseudocentre.

- In assessing the complementarity of the query-cavity surface and the ligand, it can help to focus in on just one particular part of the surface, e.g. the hydrophobic portion. The surface display is coupled to the pseudocentre display, so we do this by turning off all the pseudocentres other than those that are hydrophobic.
- Use the pseudocentre tick boxes to turn off the display of non-hydrophobic pseudocentres, viz. *Donor*, *Acceptor* and *Donor-Acceptor*. You will need to click each tick box twice. The first click will turn on all pseudocentres that are not already displayed (the tick box will change from grey to white when this happens), the second will hide all the pseudocentres of that type. You are left with just the hydrophobic pseudocentres and the corresponding hydrophobic parts of the query-cavity surface:



- It is possible to control the display of individual pseudocentres. Click on the + icon for the *Pseudocentres [Aromatic]* branch in the hit cavity. This will open the branch show all such pseudocentres in the hit cavity and their display state (Pseudocentres for Phe 330, Tyr 124 and Tyr 341 should all be shown as displayed). Individual pseudocentres can be hidden or displayed via the tick boxes in the tree.

This ends the tutorial.

4 Tutorial 4: Using the Cavity Information Module in a More In-depth Way

4.1 Overview

4.1.1 Objectives

- To find ligands that might bind to a particular protein binding site. This binding site of known interest will be used as the query cavity in a CavBase search for cavities with similar surface properties in the binding site. If a hit cavity is sufficiently similar to the query, then any ligand occupying such a cavity might also bind to the query cavity. Such ligands may be of interest in their own right or may be the source of new ideas about possible hybrids or derivatives.
- In addition, where a series is being pursued as a consequence of activity in a known active site, a CavBase search might highlight alternative binding. Such binding might possibly be competitive and thereby detrimental to the efficacy of the putative drug candidate or possibly extend the spectrum of utility.

4.1.2 Steps Required

- Prepare a hitlist so that we only search a subset of the database.
- Set up a search query to match a partial subsection of the cavity.
- Specify search options and run the search.
- View a hit cavity together with the query cavity in the 3D viewer, using various viewer options to assess how well the two match.
- Experiment with some other viewer settings.

4.1.3 The Example

- 1BXO is an aspartic endopeptidase that is complexed with a strongly binding cyclic transition state mimic inhibitor, PP7. Other examples of aspartic proteases include HIV-protease and Cathepsin D. An even closer sub-family group are the BACE proteins, which are strongly implicated in the progressive formation of insoluble amyloid plaques and vascular deposits consisting of beta-amyloid protein (beta-APP) in the brain.
  - In this tutorial, we will use CavBase to search for other ligands that might bind to this protein.
  - We will then export these ligands into a mol2 file after they, and their corresponding cavities, have been superimposed onto a common reference frame.
  - We will also show how CavBase can be used to search for and highlight movement of residues within a set of homologous structures.

4.2 Setting Up a Hitlist

- Open Relibase+.

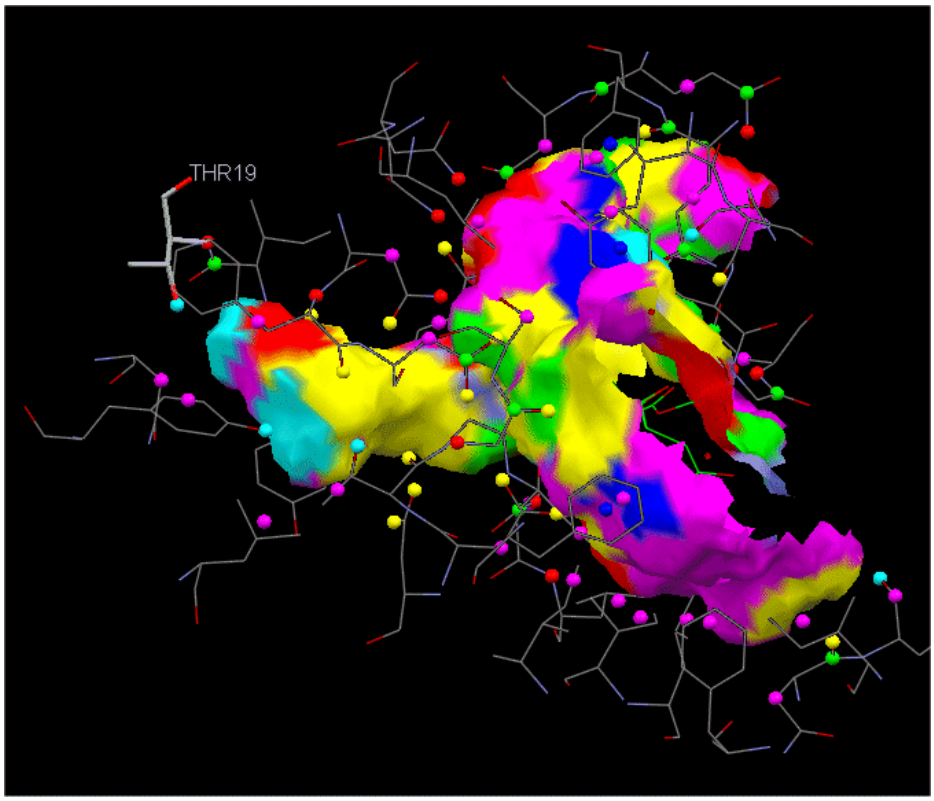
- When running a cavity similarity search, we could, if we wished, search the whole database, but that would take a long time. To speed up the tutorial we will therefore search on a small subset that we know will contain enough examples of the CavBase features we need to highlight.
- To do this we first need to create our small subset hitlist using a keyword search.
- So, hit the **Text Search** button, select *Keyword* from the *Search Type* drop-down list, and type `acid proteinase` into the *Search String* box. Because we wish to include some structures that are NOT necessarily acid proteinases, ensure the *Use regular expressions* tick box is activated.
  Note: activating the *Use regular expressions* tick box means you will retrieve structures that have comments like the following string in the header column: *...metalloproteinase with the amino-acid sequence….*
- Without changing any other settings, enter a suitable name in the *Save in Hitlist* dialogue box (e.g. *aspartic protease*) then hit the **Submit** button to run the search and save the results to a hitlist.
- Repeat the process but this time enter `aspartic protease` into the *Search String* box and give the resultant hitlist a different name.
- We now need to combine these two hitlists. Hit the top-level **Hitlists** button. In the *Hitlist Operations* frame select your first hitlist from the dropdown menu under *Protein Set 1*. Similarly select your second hitlist from the dropdown under *Protein Set 2* and then combine the two by selecting the *OR* command from the *Operation* dropdown menu. Enter a suitable name, e.g. `tutorial`, into the *New Set* box and hit the **Submit** button to create the new combined hitlist.
- When the job is done there will be a *tutorial* hitlist stored which we will use as the database subset for our cavity search. If you look for this new list in the frame containing all stored hitlists you will find a *Protein* type *tutorial* hitlist containing around 200 entries.

4.3 Setting Up the Search Query

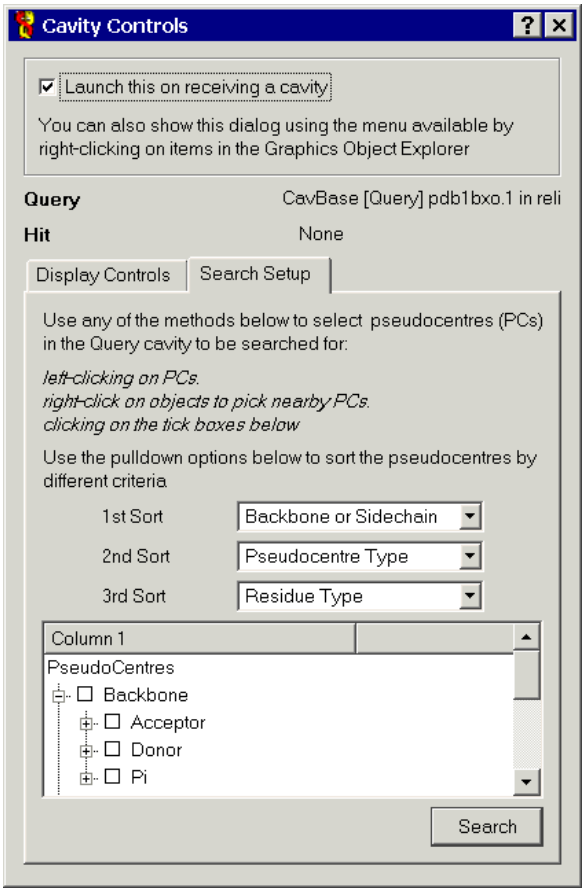
- At the top right of each Relibase+ window you'll notice a *PDB Entry Code* window:

- In the *PDB Entry Code* window, type `1BXO` and press **View**.
- Go to the very bottom of the protein information page for 1BXO and click on the **Cavity Information** button.
- This will take you to the Cavity Information page for 1BXO where you will see that just one cavity has been identified for this protein structure and entered into the CavBase database.
- Click on the link marked *pdb1bxo.1*. Some additional information about 1BXO will appear. Click on the *Load in Hermes* hyperlink to view the cavity in 3D.

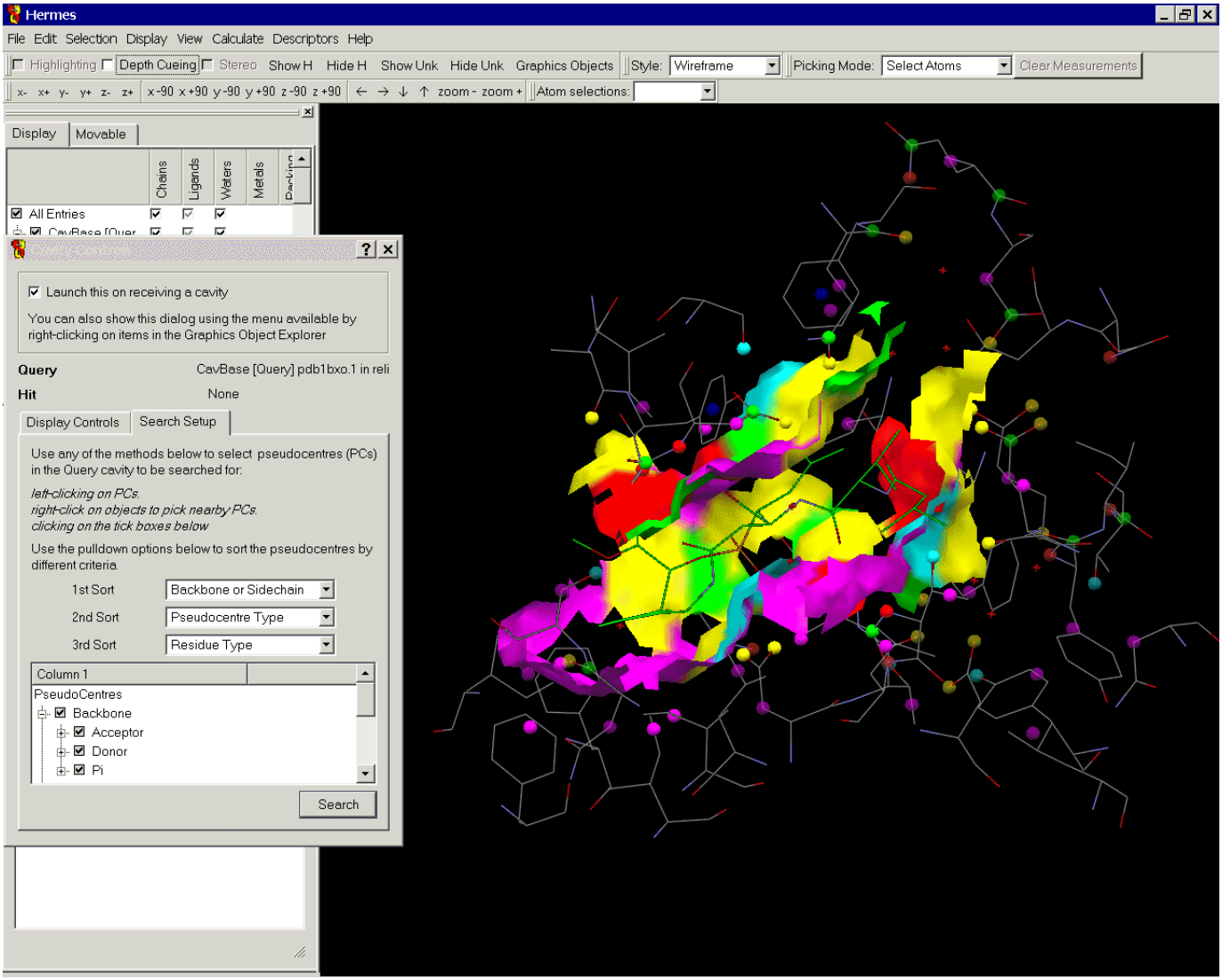
- You'll notice immediately that the amino-acids of the receptor are (a little unusually for a PDB entry) already protonated. The view can be simplified by hiding the H atoms using the **Hide H** button at the top of the 3D display.
- If you rotate the cavity in the browser you will also notice that there is a long thin extension of the cavity down towards residue THR19.



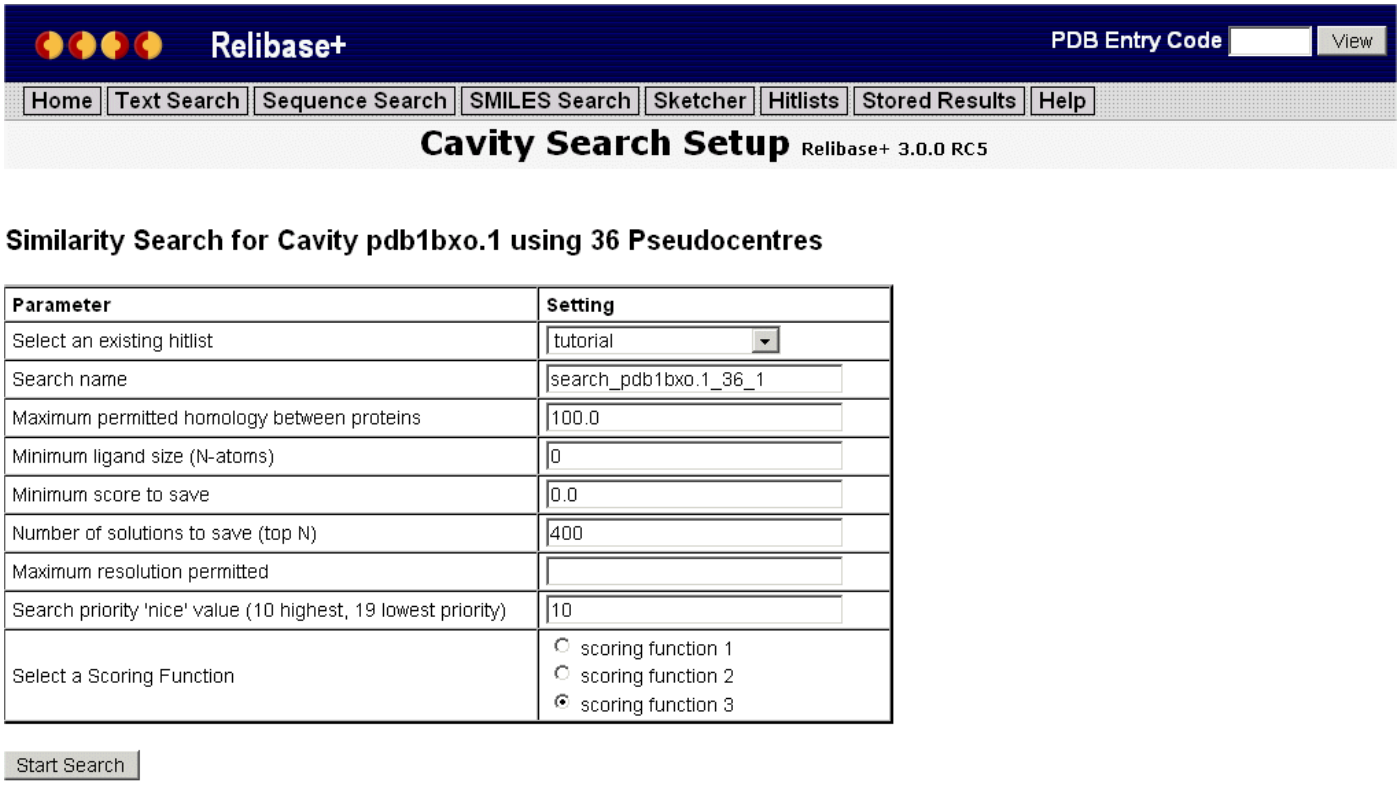
- By toggling the *Active Surfaces* off and on using the *Graphics Object Explorer* it can be seen that in this query structure this sub-region of the cavity is empty.
- A first impression is that this might be the sort of region that could accommodate a lysine or an argenine residue. Closer inspection however indicates an absence of strong H-bond acceptors in the receptor around the end of this extension that might stabilise such fragments. Indeed most ligand fragments that one might envisage filling this part of the cavity would have to be quite long and flexible and consequently not strongly bound. For the purpose of this tutorial we will remove the PseudoCentre (PC) points associated with this region of the receptor. In fact given that the pentapeptidomimetic ligand is itself quite large when compared with many pharmaceutically interesting ligands we shall reduce the full set of 90 pseudocentres identified in the cavity to just those within 4.0 Angstrom of the existing ligand in the query cavity.
- Click the **Search Setup** tab in the *Cavity Controls* window.



- You will note that the PseudoCentre points are no longer highlighted (active) and that the receptor surface associated with them is also no longer displayed.
- Right-click the mouse when the cursor is sitting on one of the ligand atoms. Two options appear: *Select PseudoCentres within range of this ligand*; *Select PseudoCentres within range of this atom*. With the left mouse button select **PseudoCentres within range of this ligand**.
- A *Select to Range* window will appear. Accept the default value of *4.00* by clicking on **OK**.
- The Hermes visualiser will resemble the following:



- At the foot of the *Cavity Controls* window press the **Search** button.
- A *Cavity Search Setup* webpage will be launched where cavity search parameters can be configured. We will discuss later some of the factors that may be considered when considering the number and type or spatial density of preferred PCs for any specific query. However for the purposes of this tutorial we will search using 36 pseudocentres.
- From the *Select an existing hitlist* pulldown menu, select the *tutorial* hitlist you saved earlier.
- Modify the *Number of solutions to save (top N)* value of *100* to *400*.
- Change the *Select a Scoring Function* radio button to *scoring function 3*.
- Keep all other settings at their defaults. The Cavity Search Setup window should resemble the following:



- Press the **Start Search** button.

4.3.1 Some Observations on Dealing with the Search Results

- After starting the cavity similarity search a new Relibase+ page will appear, the *Cavity Search Status* page, which after about one minute changes again to highlight the current status of the search and a second table that restates the search settings used.
- Three options are available at the foot of the page:
  - **Click here to view all search results**: view previous search results whose data have been saved.
  - **Click here to view current results for this query**: view current search results.
  - **Finish this query now**: stop the search. Note that search results obtained to this point will be saved and can be viewed.
- After a few further minutes, click the second option to view the current results for this query. A Search Results page will appear with data on *Score*, *Matched Centres* (number of matched pseudocentres), *RMS* (RMS distance between matched pseudocentre pairs), *Protein Homology* and *Cavity Homology* and header details about the hit protein (sorted in descending order by the score of those cavity matches compared to date). The results page is updated every 15 seconds.
- After 15-20 minutes the search will have completed.
- The top of the Search Results page should look like this:

Comparison for Query Cavity pdb1bxo.1 (Label: 'search_pdb1bxo.1_36_1')

Search settings used

maximum permitted homology between proteins	100
minimum permitted ligand size in cavity	0
minimum score to save	0
number of solutions to save	400
scoring function	3
search all entries in	the hitlist tutorial:NAM:henderson

Search Results 1-400 of 400

Select	Cavity	Score	Normalised Score	Matched Centres	RMS	Protein Homology	Cavity Homology	Header	Title
☐	pdb1bxq.1	5806.0	97.5	35	0.187	100.0	100.0	HYDROLASE	ACID PROTEINASE (PENICILLOPEPSIN) COMPLEX WITH PHOSPHONATE INHIBITOR.
☐	pdb1ppl.1	5141.7	86.4	33	0.569	100.0	78.0	HYDROLASE(ACID PROTEINASE)	CRYSTALLOGRAPHIC ANALYSIS OF TRANSITION-STATE MIMICS BOUND TO PENICILLOPEPSIN: PHOSPHORUS-CONTAINING PEPTIDE ANALOGUES
☐	pdb2web.1	5054.7	84.9	32	0.565	100.0	90.2	HYDROLASE	ACID PROTEINASE (PENICILLOPEPSIN) (E.C.3.4.23.20) COMPLEX WITH PHOSPHONATE INHIBITOR: METHYL(2S)-[1-(((N-FORMYL)-L VALYL)AMINO-2-(2-NAPHTHYL)ETHYL)HYDROXYPHOSPHINYLOXY]-3 PHENYLPROPANOATE, SODIUM SALT
☐	pdb1ppm.1	4963.5	83.4	33	0.528	100.0	90.2	HYDROLASE(ACID PROTEINASE)	CRYSTALLOGRAPHIC ANALYSIS OF TRANSITION-STATE MIMICS BOUND TO PENICILLOPEPSIN: PHOSPHORUS-CONTAINING PEPTIDE ANALOGUES
☐	pdb1ppk.1	4876.2	81.9	31	0.453	100.0	92.7	HYDROLASE(ACID PROTEINASE)	CRYSTALLOGRAPHIC ANALYSIS OF TRANSITION STATE MIMICS BOUND TO PENICILLOPEPSIN: PHOSPHOROUS-CONTAINING PEPTIDE ANALOGUES

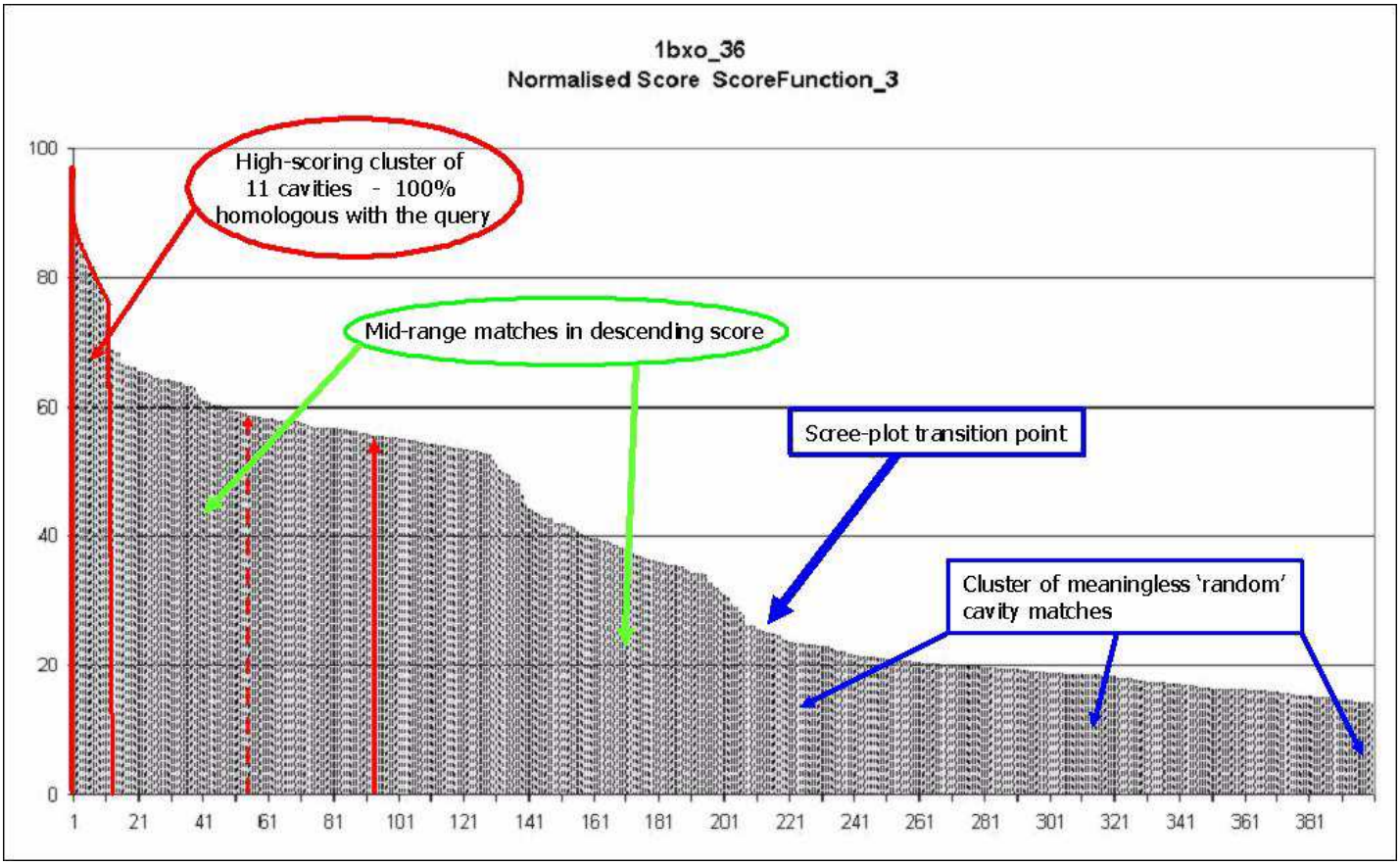
- To view the complete table of all 400 results, click **Browse all hits** (at the bottom of the table).
- One of the other options at the bottom of the page is *Download Results Table in CSV Format.* Click this option and save the complete score-ordered table as *tutorial.csv.* Return to the *Cavity Comparison Search Result* browser window.
- You will note that each matched cavity in the *Cavity* column of the table is hyperlinked to another webpage. Click the link for the first cavity in the table, **pdb1bxq.1**.
- A Cavity Information page appears, with some features of comparison between the hit cavity, *pdb1bxq.1* and the reference query cavity *pdb1bxo.1*. Click on the View in Hermes hyperlink (at the top of the results page) to view the cavities in 3D.
- Hermes will open along with a second smaller window, the *Cavity Controls* window, which you should move to one side of your screen leaving a view of the 3D display.
- You may wish to remove the query cavity's H atoms (as before) to simplify the display. It is also useful to be able to rescale these structures by moving the mouse vertically with the right mouse button depressed.
- When this is done you will notice that these two cavities are very closely matched.
- The display of protein chains, ligands and solvent can be controlled by deactivating the relevant tickboxes in the *Protein Explorer.*

- Also, the display of pseudocentre *Surfaces* can be controlled by toggling the relevant Surface boxes in the *Graphics Objects Explorer*.
- Hide the protein and solvent atoms, but keep the matched surfaces displayed.
- Now use the slider in the *Display Controls* panel of the Cavity Controls window to separate the two cavities. Given the 100% homology of the protein structures and the very close similarity between the two ligand structures it is not very surprising that the matched parts of the cavities are virtually identical. It is usually more informative to compare that part of the hit surface that matches the query with the full query surface.
- To do this, with the mouse still in the *Cavity Controls* panel, click the pull down on the *Show Pseudocentres*, *Query* line that shows pseudocentres and select *PCs searched for*. In this particular instance where 35 matches have been made out of a possible 36 there is still almost no discernable difference. The easiest way to find the single missed pseudocentre is to turn the surfaces off and then toggle between the query *matched PCs* and *PCs searched for*. The same pull-down toggle allows you to see quite how much of the initial cavity, pdb1bxo.1, has been removed from our tutorial query (*Unmatched PCs*).
- By clicking on other cavities hyperlinked in the *Search Results* table, in the Relibase+ browser, you will be able to compare other (less well) matched cavity pairs. If for example you were to compare any pair of cavities in the worst scoring 170 entries in the table you would be very hard pushed to make any meaningful interpretation.

4.4 Further optional analysis

What follows is a more detailed analysis of the search we have just carried out. The next step of this tutorial assumes you have access to Excel software although clearly there are a number of other pieces of spreadsheet software that allow the import of .csv tabulated data and its subsequent manipulation and graphical display.

- Import the *tutorial.csv* file you saved earlier into Excel (or whatever preferred spreadsheet program) and check that the *Normalised Score* column is still in descending order. It is a user preference whether or not the *Normalised Score* rather than the raw *Score* is used: *Normalised Score* gives some impression of how good a match the hit is as a percentage fit when compared with a perfect match. It also gives the impression that you can compare results when using different score functions (clearly no more accurately than you could by looking at correlations between raw scores).
- Prepare a plot of the descending *Normalised Score* for all 400 saved cavity matches.
- It should look something like this (albeit without the coloured highlighted features):



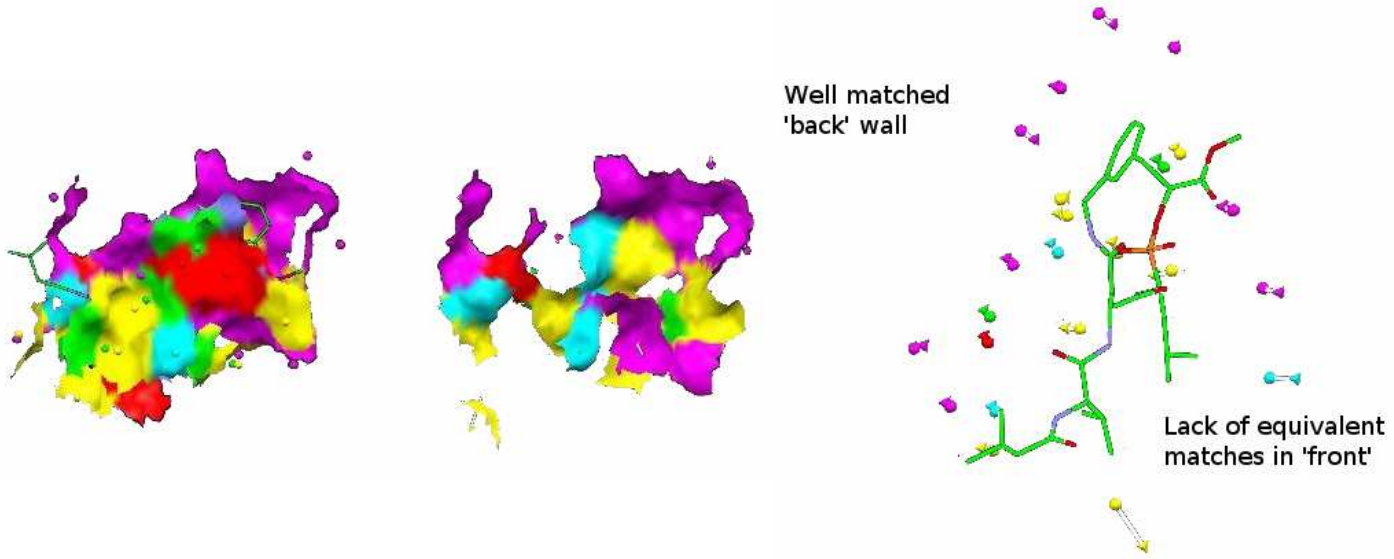
- Inspection of this plot would suggest that the 400 saved cavity matches fall broadly into three distinct groups:
  - A high-scoring cluster (Normalised Scores between 97.5-74.4%) of 11 receptors completely homologous with pdb1bxo.
  - A large group of intermediate level matches (~195 cavities with Normalised Scores between 69-27%).
  - A residual group of approximately equal size of meaningless 'random' cavity matches.
- You will also notice that in the body of the mid-range cluster are two further cavities that are components of proteins that are 100% homologous with the query (hits numbers 54, cavity and 97 with *Normalised Scores* of 59.0 and 55.3% respectively). You may find it interesting to look at both these cavities in Hermes superimposed on pdb1bxo.1 and speculate as to why their matches are so poor when compared to the 11 other homologous structures.

4.5 Identification of sidechain/backbone movement

- Return to the Cavity Comparison Results table in the Relibase+ browser and click on the hyperlink for the matched pair containing the cavity, *3app.1* (Normalised Score = 55.3). Note that if you have problems locating this entry you can order the *Cavity* (click on *Cavity*) header in the table so that the entries are ordered alphabetically/numerically. 3APP is the apo form of the

1BXO complex.

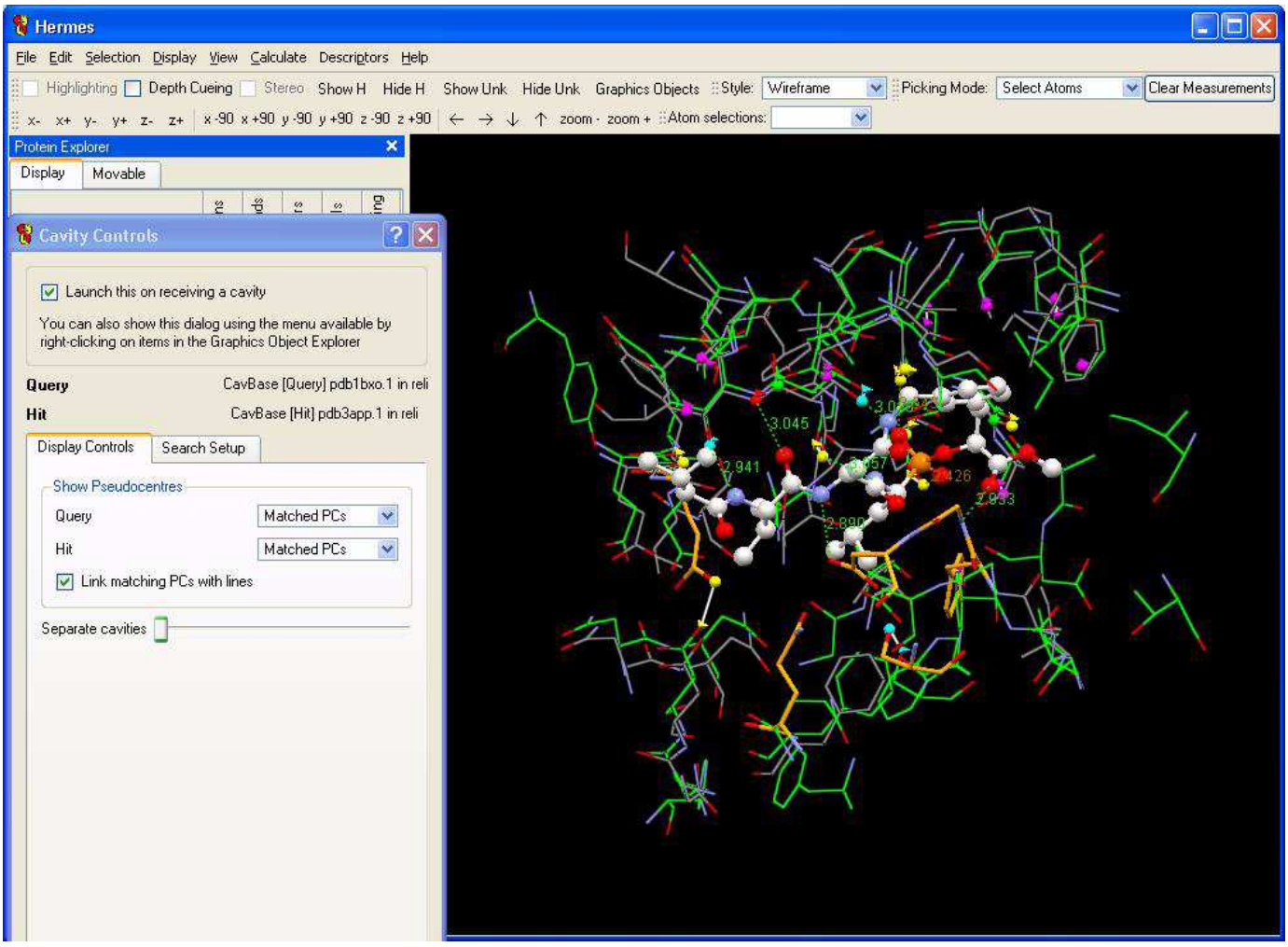
- Load the pair of cavities into Hermes as before and again toggle off the protein chains and solvent using the *Protein Explorer* panel. Make sure that the full query surface is displayed by selecting *All PCs in cavity* from the *Query* pull-down on the *Display Control* panel of the *Cavity Control* window.
- Now separate the cavities using the slider in this same panel. You will notice that although most of the 'back' surface is matched for this pair the same cannot be said for the front. You would come to the same conclusion if you slid the cavities back together again, removed all the surfaces using the *Graphics Objects Explorer*, toggled the *Cavity Controls* display to show only *Matched PCs*, then rotate the ligand.



- Display the protein chains for both cavities (H atoms undisplayed) and remove the ligand from the query protein. Gently rotating the display will make it immediately obvious that some of the amino-acid residues superimpose quite closely on top of their counterpoints whilst a few do not. Select one atom for each of those residues on the 1bxo query receptor that doesn't superpose well and then label each of those residues. You should see labels on Gly76, Asp77 and Gln111, and to a lesser extent on Tyr75 and Ser79. All of these are on the 'front wall' although you may also note a single bond rotation has taken place on Glu16 on the 'back'.
- Note these numbers and select each complete residue using the top-level menubar pull-down (**Selection**, **Define Complex Selection**, **By residue**, **Specify Individual** and then clicking on those residues - 1bxo and not 3app – before adding them to the selection by selecting **Add** and then **Close**, then **Close** again to finish the selection). The corresponding atoms will now be highlighted in the 3D view. Colour them distinctively (right-click in open space in the 3D view, select **Colours** and then e.g. **Orange**), perhaps changing style to 'Capped Sticks'.
- Inspection of the result indicates quite clearly that the inside of the receptor cavity has collapsed quite significantly on going from the apo-form of the enzyme, 3app, to the ligand bound form,

1bxo.

- The reason for such a shift is quite obvious if the 1bxo ligand is now displayed back in the receptor. You may wish to modify the display style and/or the ligand's colour for clarity.
- Highlight any H bond interactions between ligand and receptor by activating the H-bond tick box adjacent to *[Query]* in the Contacts panel of Hermes. Note that if the *Contacts* panel is not present it can be launched via **View**, **Contacts**. At least two H-bonds have obviously been formed as a direct consequence of this cavity collapse in the apo structure, quite apart from a much reduced surface area exposed to solvent. In fact if the default limits of the H-bond contact are increased very slightly (via the **Define H-bonds** button), more additional H-bonds will be evident.

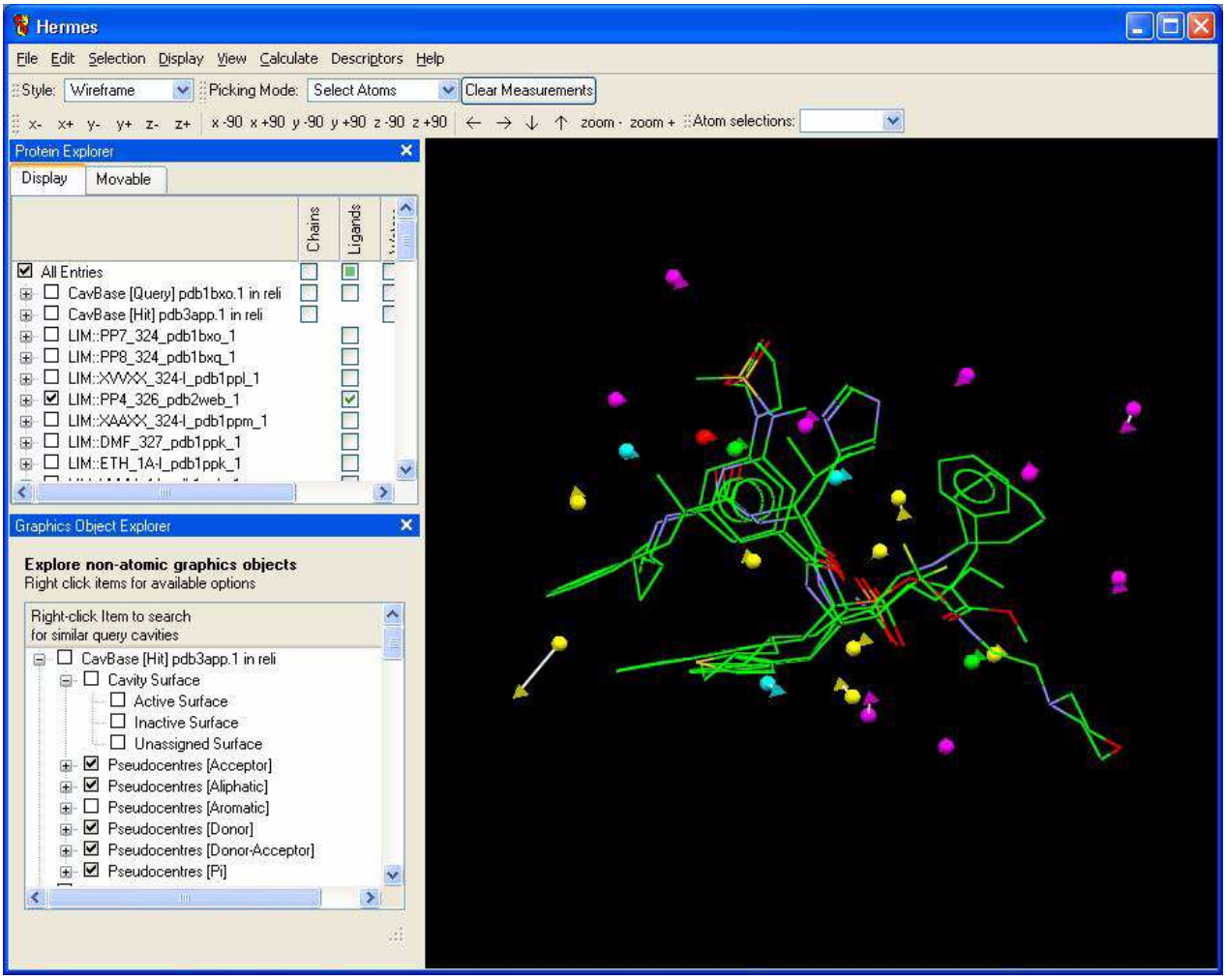

Hermes
File Edit Selection Display View Calculate Descriptors Help
Highlighting Depth Cueing Stereo Show H Hide H Show Unk Hide Unk Graphics Objects Style: Wireframe Picking Mode: Select Atoms Clear Measurements
Protein Explorer
Display Movable
Cavity Controls
Launch this on receiving a cavity
You can also show this dialog using the menu available by right-clicking on items in the Graphics Object Explorer
Query CavBase [Query] pdb1bxo.1 in reli
Hit CavBase [Hit] pdb3app.1 in reli
Display Controls Search Setup
Show Pseudocentres
Query Matched PCs
Hit Matched PCs
Link matching PCs with lines
Separate cavities
3.045
2.941
2.933

- As a side issue, you will notice that inspection of H-bond contacts highlighted in this way on the back wall indicate apparently two quite short H-bond contacts between the phosphonate fragment of the ligand and the acid residues Asp33 and Asp213, where one would not intuitively expect to see such an interaction. One might speculate about the unseen presence of a Mg ion or other similar small cation. You could even treat it as an excuse for another Relibase 2D-sketcher search and analysis!

- It may be interesting to repeat the same exercise with the other low-scoring 100% homologous protein structure, 2wea. Since this PDB entry contains a ligand structure very similar to that in 1bxo, the real question might be to explain why the lack of collapse.
- It might also be useful to look at the superimposition of the cavity 3cms.2 on our query cavity in the same way. This protein, which is only 30.8% homologous with our query structure and only a 39% cavity match (17 PCs matched out of 36) looks quite remarkably similar in some aspects. Why? Where might one suspect the mutational changes have been made?

4.6 Inspection of superimposed ligands

- When the initial CavBase search had been completed you may have noted that at the foot of the *Cavity Comparison Search Results* page were some additional options.
- One of these is **Superpose Selected Cavity Binding Sites**. In fact the default state is that all the structures are superimposed which is why you didn't have to use this option before running the previous part of the tutorial. However, on occasion you may wish to look at specific multiple superimposed cavities in Hermes without having to load the entire set of solutions (which in this instance would make Hermes run slowly).
- To superimpose selected hits we must first select the structures of interest by activating the tick box in the first column of the table. Select 2oah (2 cavities), 2web (1 cavity), 2vj9 (3 cavities) and 1epr (1 cavity), then click on **Superpose Selected Cavity Binding Sites**.
- To view the superimposed structures that we have selected, click on the **Display Superimposed Cavity Binding Sites in Hermes** hyperlink. This option offers you the possibility of inspecting the various binding sites and ligands after they have been superimposed according to a receptor based rationale.
- There are a number of existing methods for superimposing ligands but not many that deal reasonably with dissimilar structures.
- One of the objectives one might wish to accomplish with such an aligned set of structures is to use the set as a source of ideas for de-novo hybrid compounds.
- It is unfortunate that the subset of protein structures saved from your original cavity search contained a large number of meaningless matches. Coincidentally a large number of those matches that are significant contain oligopeptide-like ligands, which do not particularly lend themselves to the formation of drug-like novel hybrids. A realistic tutorial trying to demonstrate this approach requires a new cavity similarity search.

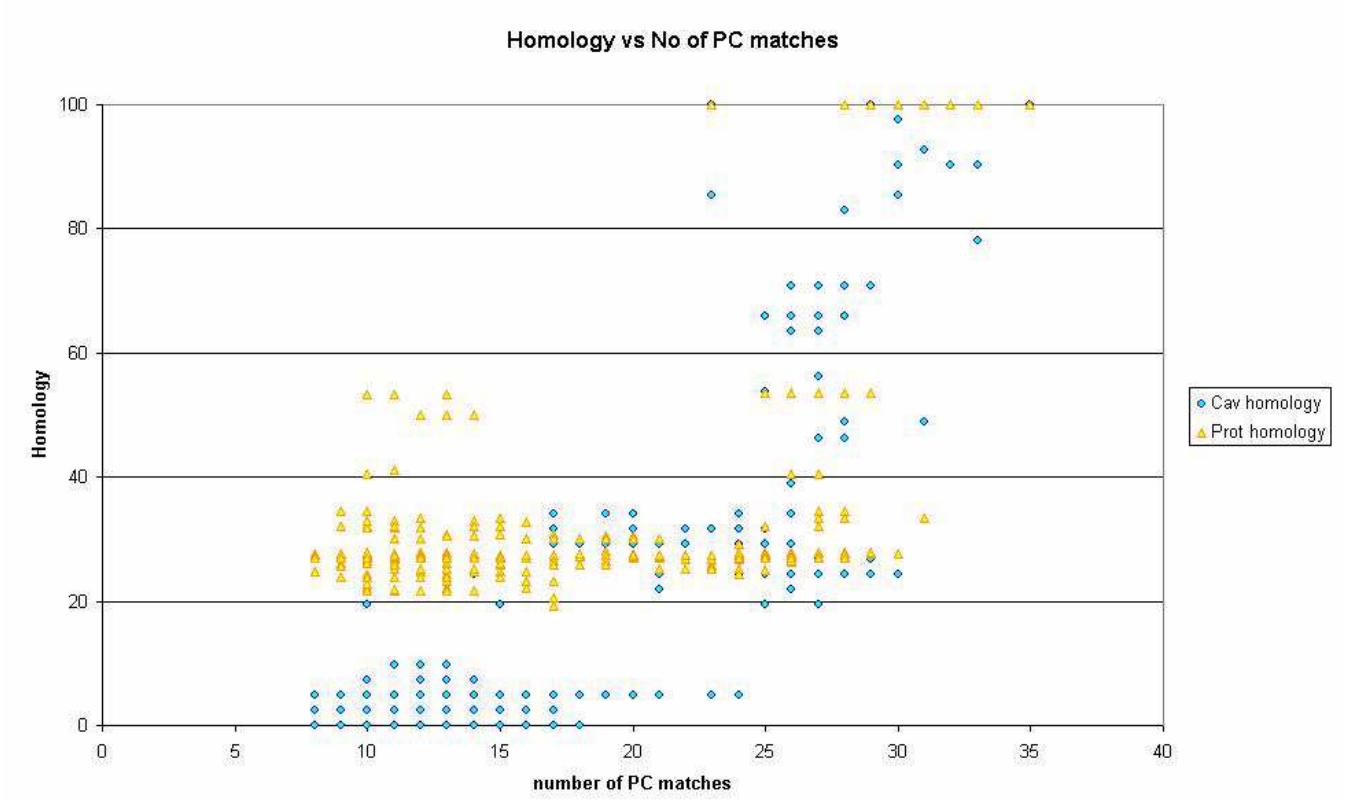


- To try and make it a drug-like structure, bearing in mind the sorts of parameters that would influence bioavailability or metabolic stability, is of course a little more problematic.
- If you look at the superimposed ligands from 2oah, 2web, 2vj9 and 1epr, you may get some idea of how such a set might help produce ideas for a non-peptide penicillopepsin inhibitor.

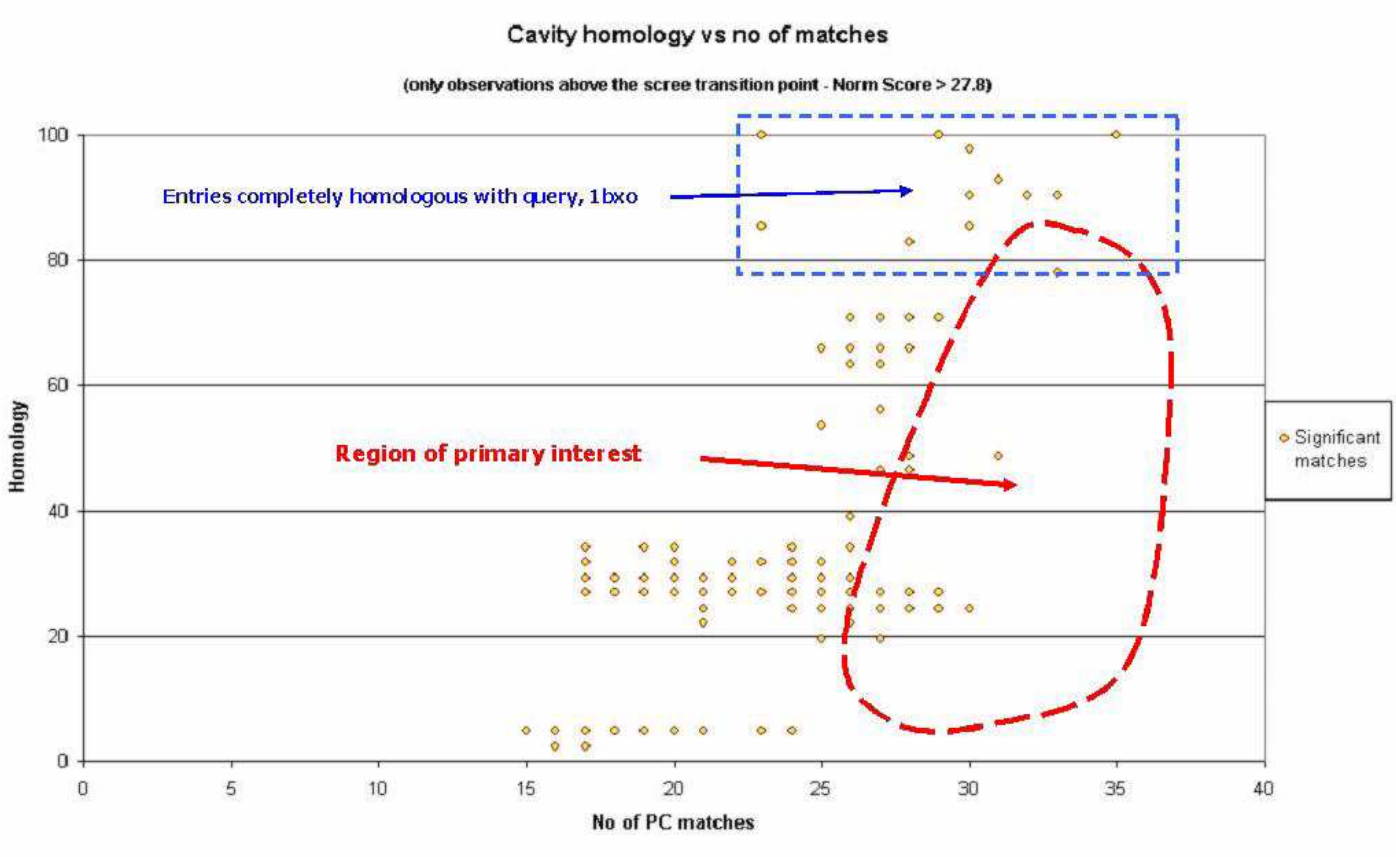
4.7 Pairwise comparison between cavity/protein homologies and no. of PC matches

- It is sometimes interesting to compare these matched homologies with each other and with their corresponding scores and no._of_PC_matches. For example, it could be useful to identify receptors with little overall sequence identity to the known query but which are highly homologous in those residues around the cavity and perhaps therefore functionally equivalent (although this sort of information is likely to be already known using other methodologies). More realistically it would be useful to identify structures with little direct cavity homology but a significant number of PC matches, structures that might increase the chance of finding matched cavities that do not have already-known equivalent functionality, where bound ligand overlays may provide novel insights towards de-novo design.

- Given that the data-set searched in this tutorial was preselected on the basis of known functionality, aspartic proteases and acid proteinases, this search you have just conducted is unlikely to provide a good example of such a match. However, if you return to the spreadsheet file *tutorial.csv* that you prepared earlier and plot both the protein homology vs number of PC matches and also the cavity homology against the same x-axis you should obtain something like the plot below. Clearly, apart from the obvious existence of a cluster of entries 100% homologous with 1bxo, nothing much can be inferred from the protein homology data but the cavity data does reflect the equally obvious correlation between cavity homology and PC matches with the query active site. You should remember that for this specific CavBase search you have already demonstrated that about half the pairs that were saved (those with scores below the scree transition point, *Normalised Score* ~ 27.8) are not meaningful matches.

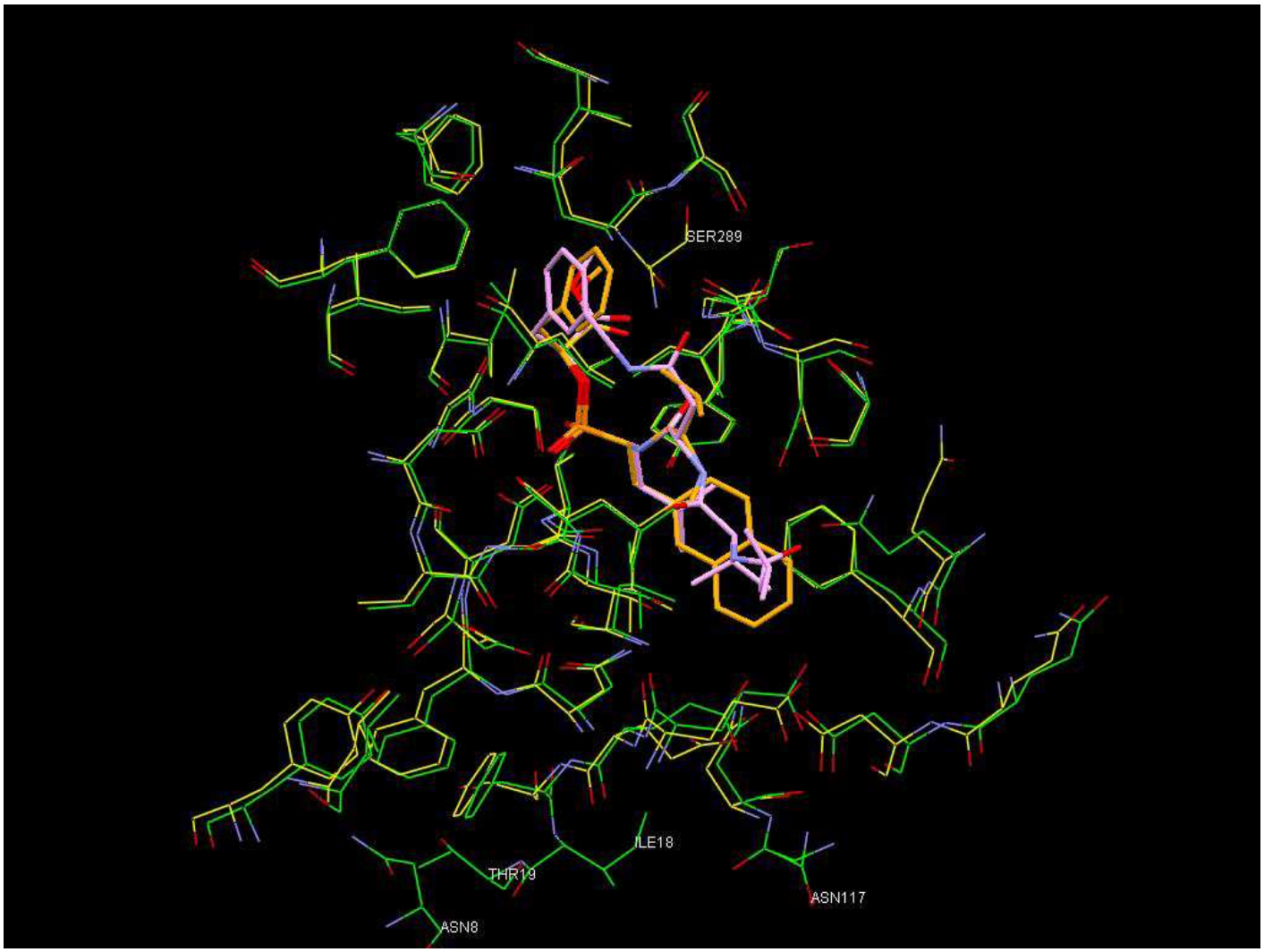


- Bearing in mind this last observation, the cavity data are replotted below with these points removed and a couple of other features have been highlighted.
- It might seem initially worrying that cavities extracted from structures 100% homologous with the query protein are not necessarily 100% homologous themselves with the query cavity.



- But the Ligsite software used to identify and extract the ligand-free cavities may find that even though the full proteins may be completely homologous, when entries containing different ligands (occasionally even identical ligands bound to different chains in the same oligomeric entry!) are measured, the resulting cavities are not always bounded by the same number (or identity) of residues. Thus in the example shown at the end of this section, the cavity extracted from the penicillopepsin, 2web, does not contain the residues ASN8, ILE18, THR19 or ASN117 although it does include SER289 unlike the query site and in this case the CavBase cavity homology calculation will record the absence of these four residue matches as a reduction of the 100% homology noted for the complete protein chains to 90%.
- Because the initial search was limited to structures with a known strong functional similarity to the query receptor it is not surprising that the region of primary interest, that part of the plot where there is low cavity homology but high PC matching, is disappointingly vacant.
- Suffice to say that most of the hits found in this region of the plot above are BACE enzymes, which themselves are part of a cluster of strongly conserved structures with a well-known partial similarity to the penicillopepsins.

4.8 Superimposed CavBase cavities obtained from 1bxo and 2web



- Although these these two cavities superimpose very closely, RMS 0.56 Angstroms, you will notice the four extra labelled amino-acids in the 40-residue query 1bxo (green) at the bottom of the figure which are not matched in the corresponding 2web hit (yellow) and which account for the resultant cavity homology of just 90%.

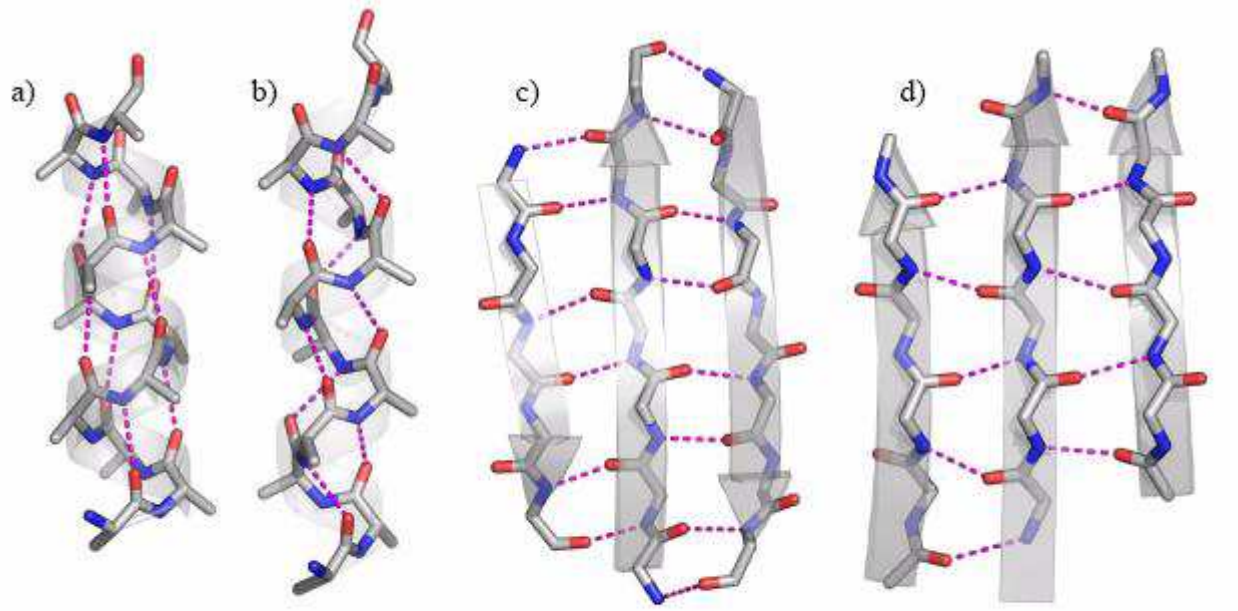
5 Tutorial 5: Introduction to the Secondary Structure Module

5.1 Objectives

- To use the protein secondary structure database, SecBase, its turn-classification and its SHAFT helix-assignment H-bonding classification to probe possible associations with turn elements and special modes of kinase inhibition.

5.2 SHAFT classification - recap

- Relibase+ has historically stored structural files in a .pdb format that in addition to the atomic coordinates and residual sequence, has appended a certain amount of secondary structure associated with each protein entry. This secondary structure has been automatically assigned by software using methodology closely related to the DSSP algorithm of Kabsch and Sander. DSSP recognizes eight types of secondary structure depending on the pattern of hydrogen bonds. The 3_{10} helix, alpha helix and pi helix are recognized by having a repetitive sequence of hydrogen bonds in which the donor residue is three, four, or five residues later in the backbone.
- In addition DSSP recognizes two types of hydrogen-bond pairs in beta sheet structures, the parallel and antiparallel bridge.



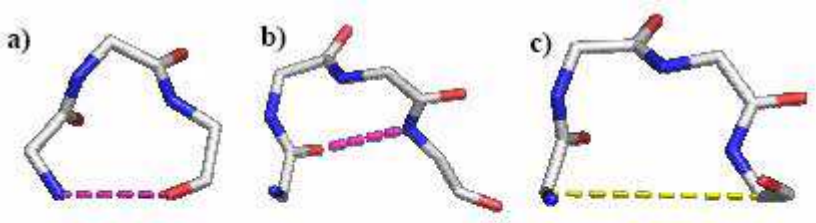
Schematic view on the hydrogen-bonding network along a right handed α helix (a), 3_{10} helix (b), antiparallel β sheet (c) and parallel β sheet

- Whilst these are the principle features recognised by the DSSP algorithm, the procedure does recognise two additional turns in terms of H-bonds but most commonly leaves other turns blank when no other rule pertains.
- The SHAFT classification scheme is a new scheme, published soon (SHAFT: Secondary

Structure - Helix Assignment From Turns, O. Koch, *to be submitted*). It is based on the recently published turn classification (O. Koch, G. Klebe, *Proteins*, **74**, 353-67, 2009) and uses a rather more consistent set of rules with regard to the termini of helical structures. The turn classification required all the turns found in a non-redundant dataset of 1903 protein chains to be clearly defined in terms of specific characteristics; automatically clustered and then systematically classified. One result of such an extensive classification is that not only are the commonly recognised helical and sheet structures identified but that the bulk of the remaining secondary structure, much of which was hitherto unclassified, is also classified in an objective systematic manner in terms of a relatively small number of turn types. This includes a large number of interesting protein regions that previously could not be treated in this way (eg. catalytic serine protease triad, DFG regions of tyrosine kinases etc).

5.3 Turn Classification - recap

- As already mentioned in the previous section, the recently published turn classification (O. Koch, G. Klebe, *Proteins*, **74**, 353-67, 2009) is based on a non-redundant dataset of 1903 protein chains. The definition of the turn family is firstly based on a hydrogen bond between CO_i and NH_i+n and then, where there isn't an internal H-bond, on the Cα-Cα distance subject to a distance constraint of less than 10Å. During the analysis following on from these definitions three different subcategories for turns based on the hydrogen-bonding pattern between the first and the last residue have been introduced (see the figure below):
  a) A reverse conformation with a hydrogen bond between NH_i and CO_i+n
  b) A standard or normal conformation with a hydrogen bond between CO_i and NH_i+n.
  c) A distorted or open turn conformation lacking a hydrogen bond, with a $Cα_i$-$Cα_i$+n distance < 10 Å.

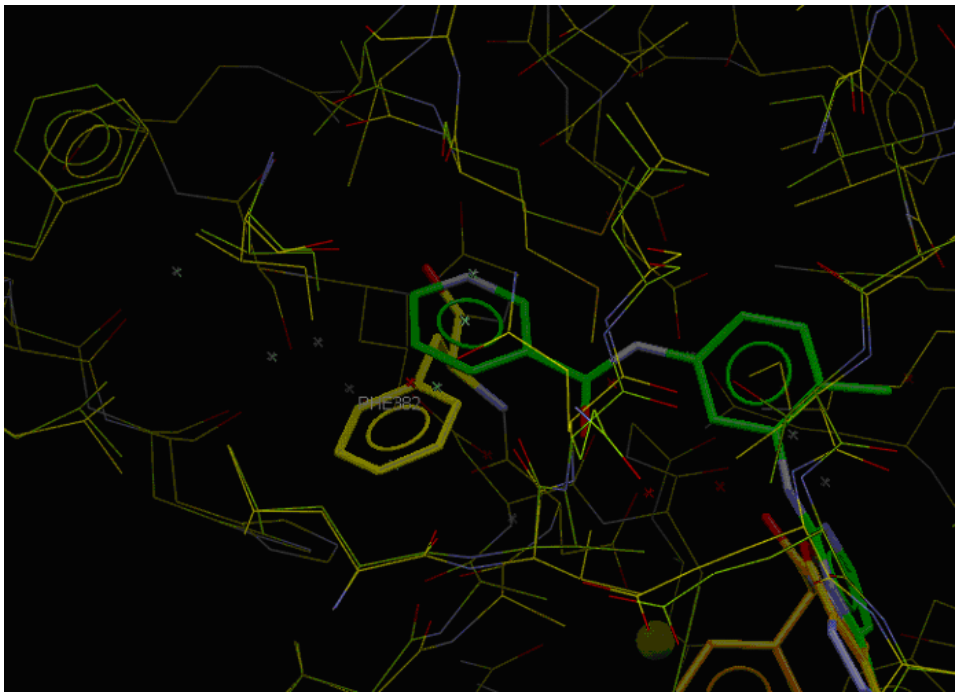


Different turn conformation
a) reverse turn with 3 residues, b) normal β-turn (4 residues),
c) open β-turn (4 residues)
magenta: hydrogen bonding, yellow: Cα-Cα distance

- The inner residues of a normal or open turn are those turns that lie within the described hydrogen bonded ring. The φ and ψ angles of these residues and the additional L angles are used for clustering.

5.4 The Example:

- In 2000 the structure of a new kinase inhibitor was published, which when crystallised in its receptor complex 1fpu (parent enzyme: ABL Kinase), was shown to have a distinctly different binding mode from those kinase inhibitors known previously. Hitherto inhibitors were found to act by direct competitive replacement of a unique essential substrate (ATP) at a very highly conserved site and with an equally conserved set of interactions. This mimicry of highly polar ATP would clearly lead to problems of inadequate specificity and often a difficulty in being able to make new inhibitors as lipophilic as other pharmaceutical considerations might prefer.
- The 1fpu X-ray showed that this inhibitor occupies a new allosteric binding pocket spatially distinct from the ATP/hinge region. The image below shows a Relibase+ superposition of the 1fpu structure with an identical protein structure that has not undergone this conformational change. Favourable interactions with the ligand provoke a large conformational change for residues in a section of sequence that is generally conserved over kinases. This conserved section contains a Asp-Phe-Gly chain known as the DFG loop, and the Phe residue, previously buried in what now becomes a strongly lipophilic pocket, moves by ~10Å to a position where the side chain occupies the space required by the phosphate groups of ATP.



- In this example we use the ReliBase+ cavity information module to identify the secondary structure characteristics of this rearranged *DFG-out* region using the SHAFT protein classification. We will also combine this with a 3D search to simply and quickly search the Relibase+ database for other examples of this *DFG-out* conformation. This is the sort of job one might wish to do when attempting to set up a pharmacophore for new inhibitors of this allosteric

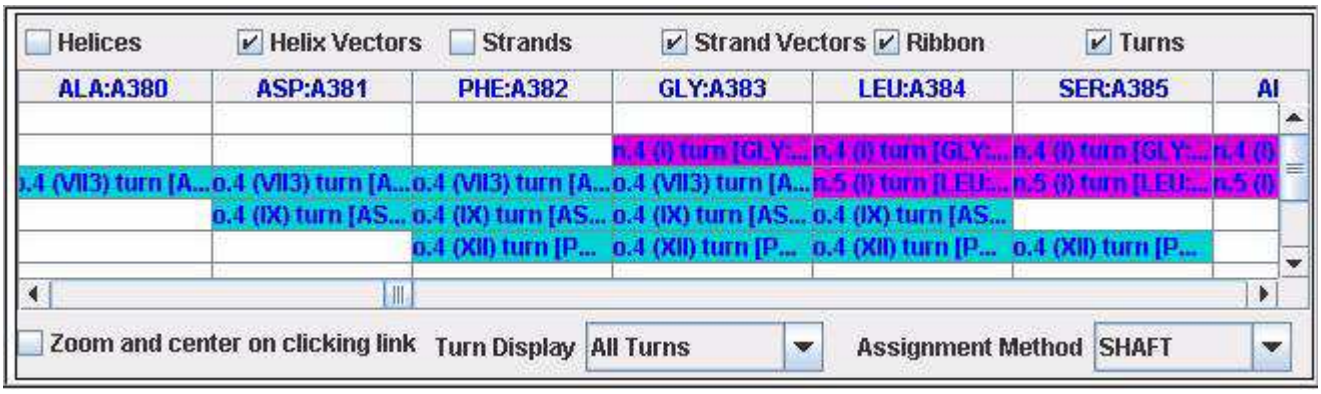
site.

5.5 Secondary structure classification of the DFG-region

- Open Relibase+.
- In the *PDB Entry Code* window on the top right of the Relibase+ interface, enter `1fpu`. The *Protein Information* page for the complex will appear. The structure is a dimeric structure.
- At the foot of the page is a table entry labelled **Secondary Structure Information**. Click on this button.
- A rotatable 3D-image of the dimer will be displayed in the embedded 3D visualiser with some of its dominant features highlighted in a ribbon format such as regions of helices (red), sheets (yellow) and remaining turns (blue). In the *Assignment Method* pulldown menu at the bottom of the page you will see that this display is the assignment made in the original PDB file.
- Underneath the 3D-display are a set of buttons that toggle specific features on or off. Below that is a table based on individual residues for the whole of the complex that contains the information about helices and the turn type information for each of these residues. In order to display this for all the complex there is a slider at the foot of the table that enables the full table to be scrolled.
- We might start by looking to see if there is any simple qualitative difference in the assignment of helical or b-sheet structure between the original PDB assessment and the SHAFT one. In the interest of visual clarity use the toggle buttons to switch off the protein *Chains* and also the *Helix Vectors* and *Strand Vectors*. The inhibitor can now be seen firmly wedged between the N-terminal domain and the C-block.
- If you switch a few times between the options offered in the **Assignment Method** drop-down menu you will note that although the b-sheet assignment is the same for both methods in this instance, there are some small differences in the helix assessments. It may be helpful to note that the display can be zoomed by using the keyboard **Shift** button with the left or right mouse button and moving the mouse in or out. Similarly, a translational movement of the display can be made pressing the keyboard **Control** button and moving the mouse with left or right mouse button depressed.
- The same effect can be seen more easily if the *Helices* and *Strands* ribbons are replaced by their *Helix Vectors* and *Strand Vectors*.
- For this 1fpu structure the residue numbers for the relevant DFG region are Asp381, Phe 382 and Gly383.

- Before inspecting this region first redisplay the *Helix Vectors* and *Strand Vectors* but with the protein *Chains* still off.
- Adjust the slider at the bottom of the table until the columns for residues ASP:A381, PHE:A382 and GLY:A383 are visible. Initially toggle off the *Zoom and center on clicking link* tick box at the bottom left.

- Click the cell in the table marked *PHE:A382* and look at the display in the general vicinity of the ligand pyrimidine ring. If there is no phenylalanine side chain visible try the active site in the other member of the dimer! To do this you will need to translate and zoom the display using the keyboard **Shift** and **Control** keys.
- You will already have noted that under the column marking the Phe382 residue classification are a number of rows, three of which have been coloured and marked with different turn types.



- These rows indicate that the specific geometric features experimentally observed with this residue come within the defining tolerances for three different turn types. A certain amount of prior knowledge is therefore helpful when it comes to deciding which of these three turn types is most useful when it comes to solving our present problem. (Note: the turn class describes all turns of similar length, the turn-type is the specific clustered turn geometry, see the turn classification paper). But we do know that this DFG sequence is totally conserved for all the Tyr/Ser/Thr kinases that are reported to be susceptible to this conformational switch and it might be reasonable to expect such a stable configuration to be part of a single structural fragment and therefore common enough to have been identified elsewhere. We can also see from inspection of the 3D-image that the next five residues form a well ordered helical substructure which again we might expect to see with a single known classification.
- The only classification for Phe382 consistent with this information is as member of an open 4-residue type VII3 turn according to the turn classification. The shorthand version for this provided in the table cell: *o.4.(VII3) turn*.
- Click the cyan coloured cell marked as this type under the *Phe:A382* heading. The side chains will appear highlighted in ball and stick format.
- It might perhaps be of interest to repeat this process in a different browser window with a conventional ATP binding site inhibitor complex, for example the structure of the tyrosine kinase ACK1, 1u54, and compare the similarity and differences.

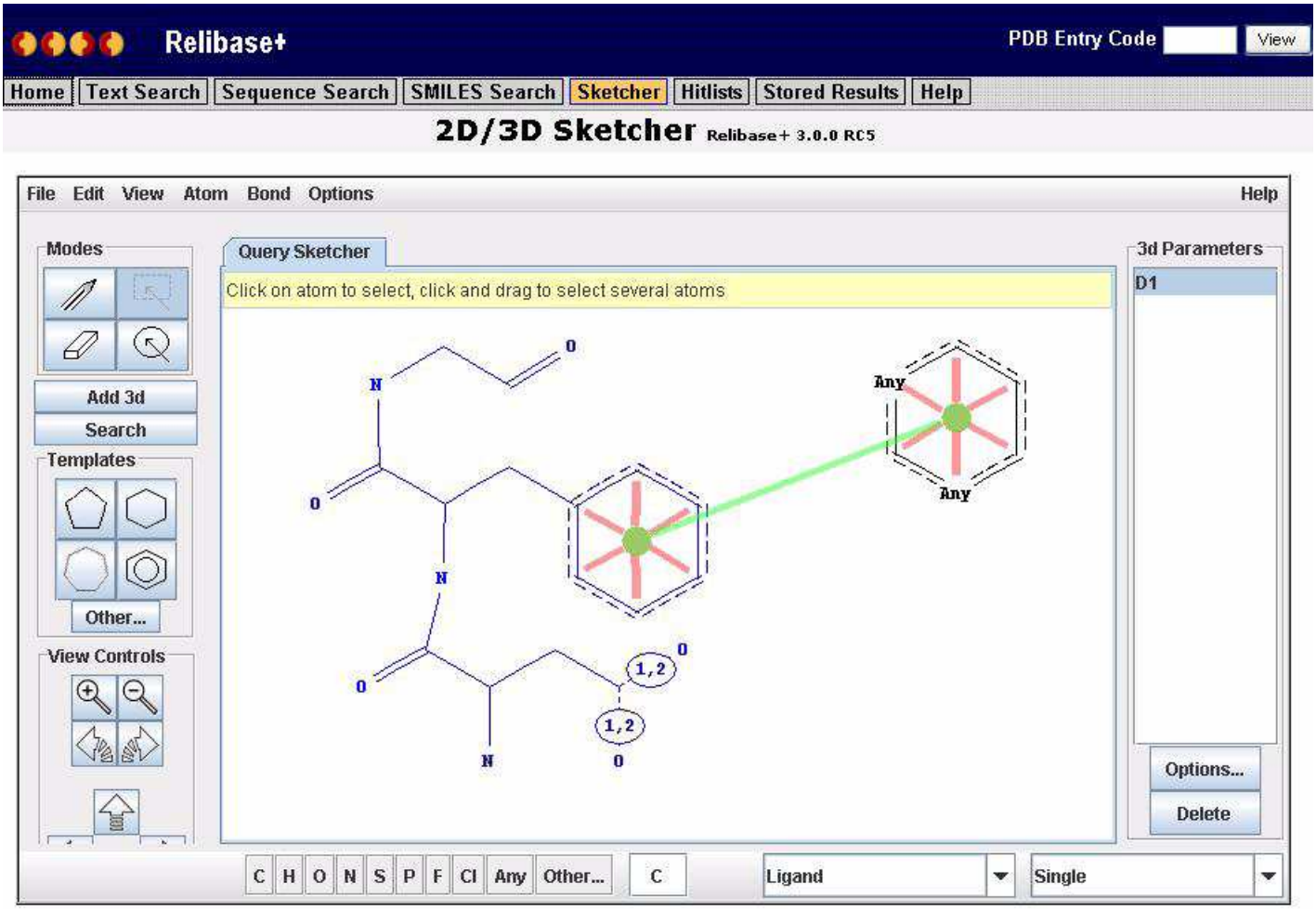
- A quick initial inspection of the ligand-filled active site of 1fpu suggests that the DFG-out protein is in part stabilised by good overlap with a neighbouring aromatic ring in the inhibitor. It might therefore help to find similar DFG conformations if we add a simple additional distance

requirement to our search and look for a ligand with an aromatic ring close to that of the Phe382.

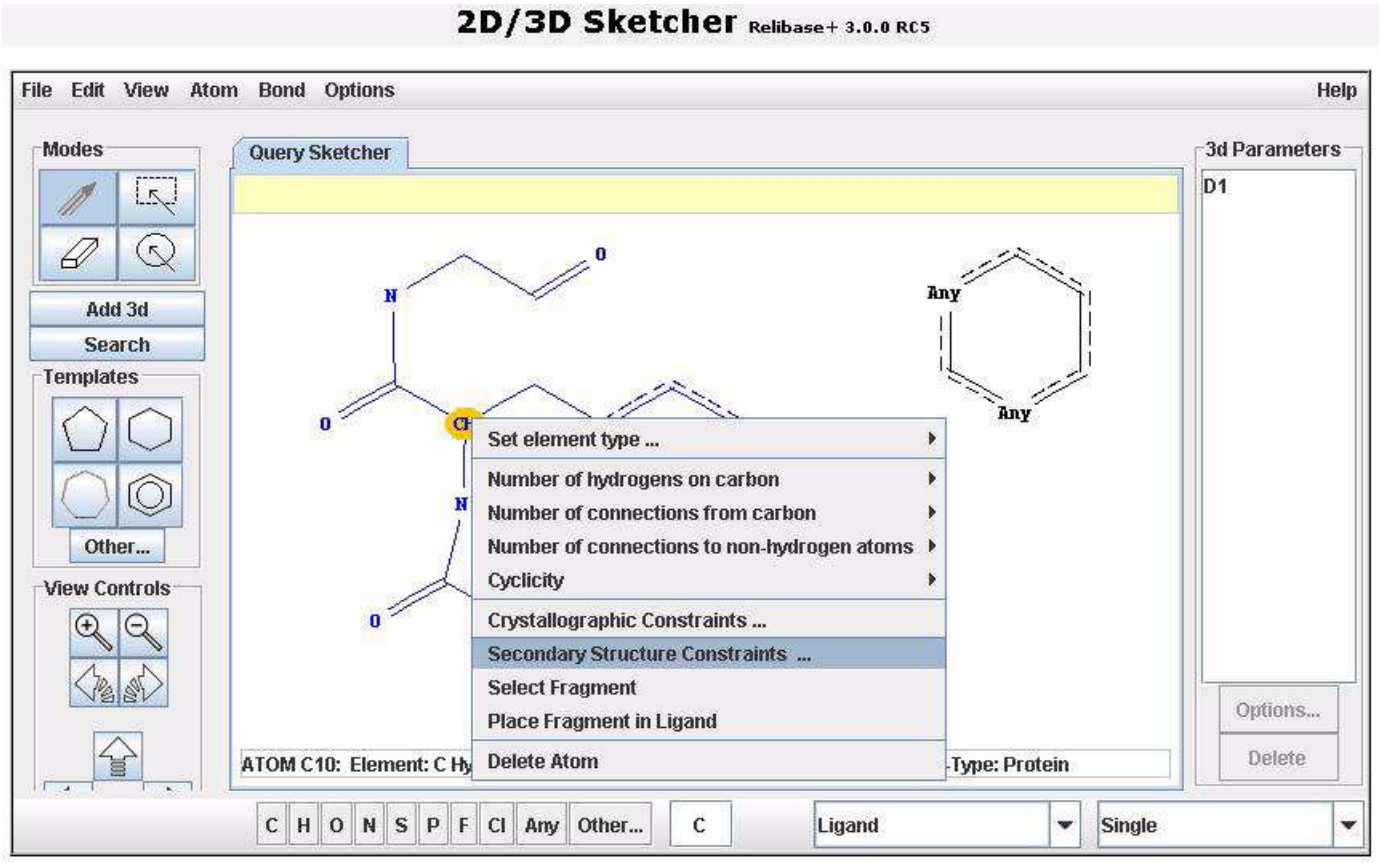
- However, a closer look at the complex could lead to many other conclusions about additional 3D constraints that would probably be characteristic of this allosteric site - for example a lipophilic ligand portion in the region vacated by the Phe382 ring would probably be as good an identifier as the one we have selected.
- And lastly, because the DFG-out pocket fillers are all kinase inhibitors (and we want this search to be quick) we will limit the search to known kinase structures.
  - Prepare the kinase subset by clicking on the Relibase+ menu option **Text Search**.
  - Select *Keyword* from the *Search Type* pulldown menu and type `kinase` into the *Search String* box.
  - Type `key_kinase` into the *Save in Hitlist* box, select *reli* in the *Use Databases* box then hit the **Submit** button to start the search. A hitlist called *key_kinase* will be saved.

5.6 Building the 3D-search query

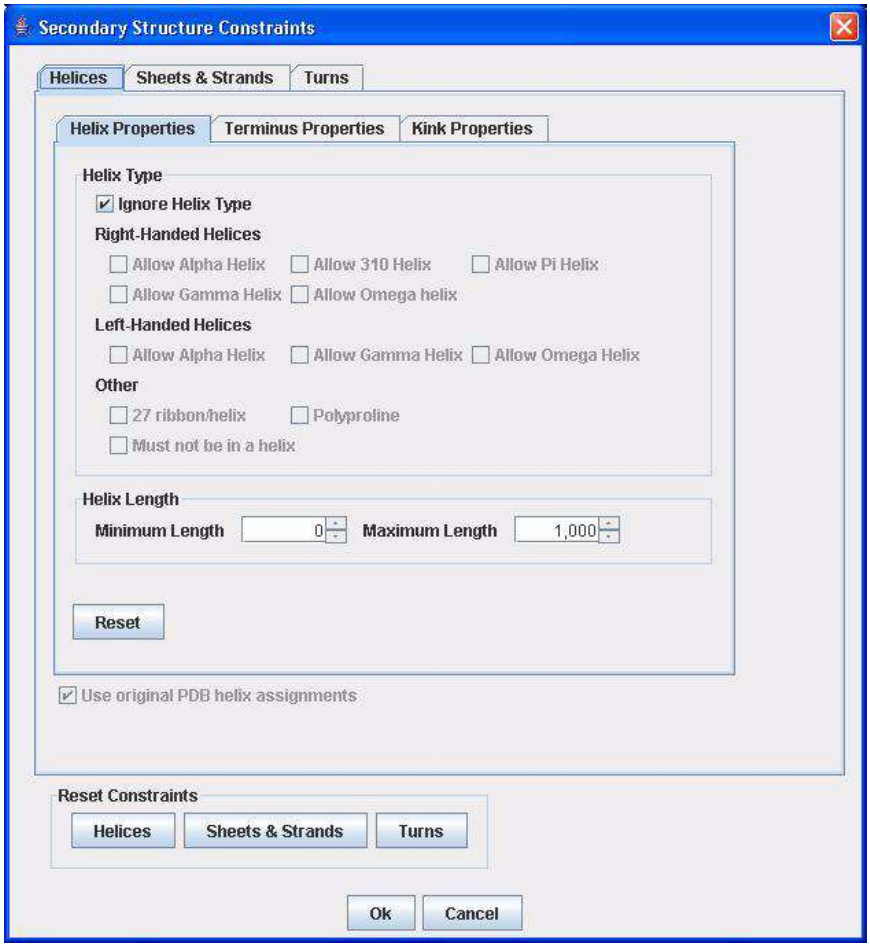
- Launch the sketcher by clicking on the top level **Sketcher** button.
- When the sketcher appears build the query until the page looks as below. Note that we have assumed familiarity with the sketcher and the construction and definition of 3D constraints such as centroids and distance ranges. Also the atomic distinction between protein and ligand atoms. Please refer to earlier tutorials or to the Relibase+ documentation for further information.



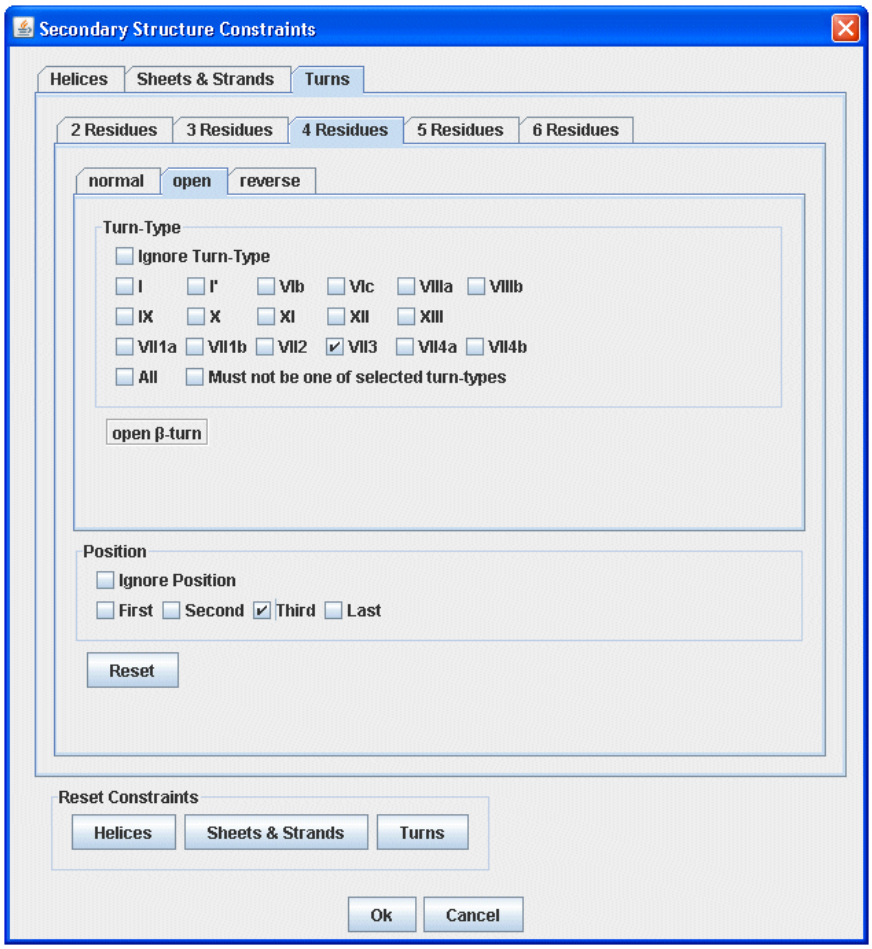
- Within the *Edit 3D Parameters* window (launched when the **Add 3D** button is clicked), define the ring centroids then define the distance between them, constraining them to a lower limit of 1.0Å and to an upper limit of 6.0Å.
- Make sure that the atoms of the DFG fragment have been typed as protein (blue) and those of the other aromatic ring as ligand type (black).



- With the cursor on any one of the phenylalanine atoms (see above) click the right-hand mouse button and from the resulting drop-down menu select **Secondary Structure Constraints**. The following panel will appear:

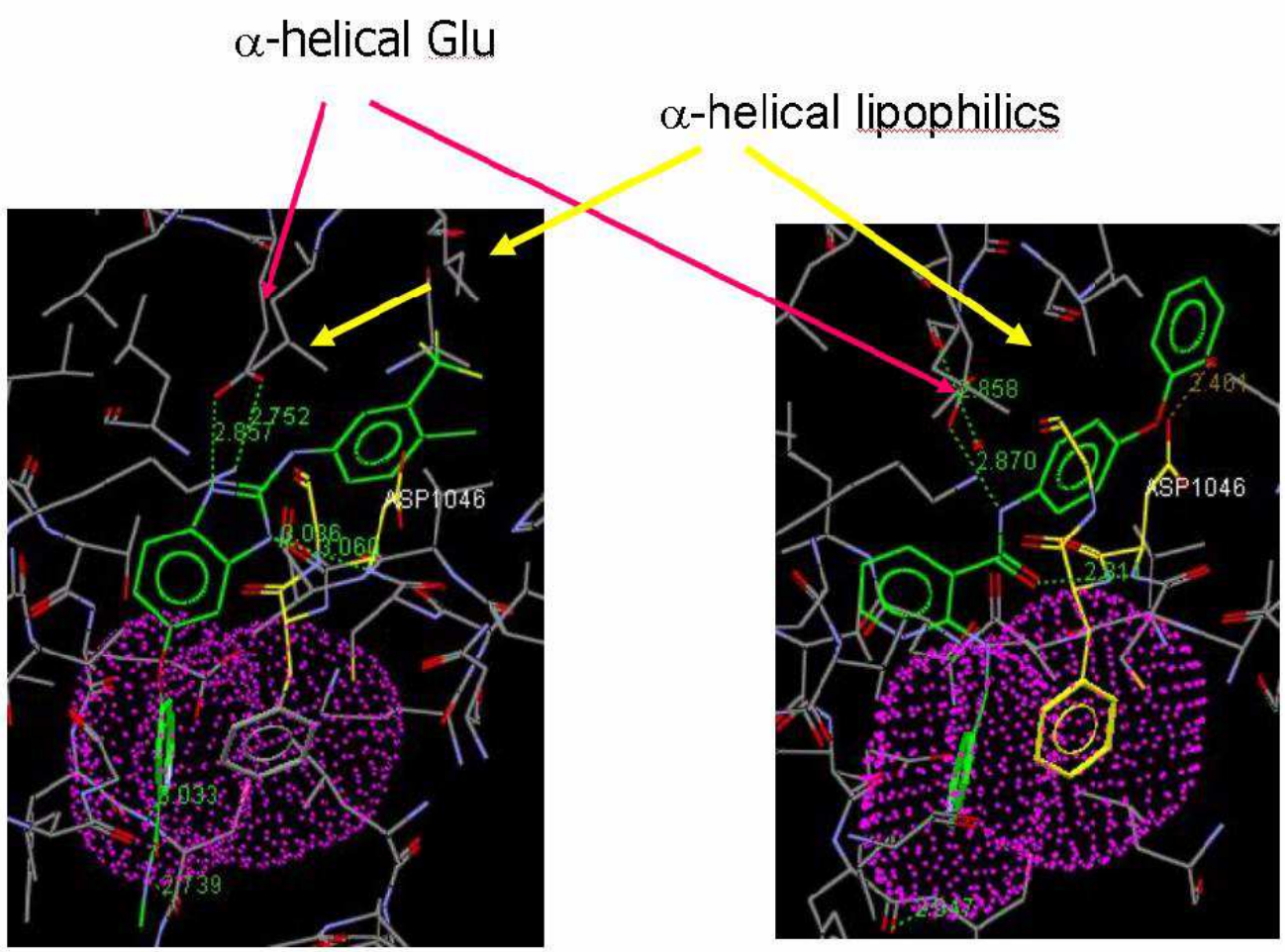


- You have already identified that the DFG substructure you wish to find is not a helix but rather part of an open 4-residue type (VII3) turn and so the default window shown above is not appropriate.
- Click on the **Turns** tab and then select the **4 Residues** tab. Now select the **open** tab and toggle off the *Ignore Turn-Type* check box.
- Activate the tick box adjacent to turn-type **VII3**.
- Because our DFG substructure is so heavily conserved (although not in fact the initial Ala residue) it is likely that the position of each residue in the fragment is also important and you should therefore toggle off the *Ignore Position* check box and check on the *Third* position.
- The *Secondary Structure Constraints* window should now look as below:



- Press **Ok** to return to the sketcher.
- Click the **Search** button to the left of the sketcher window to start the search.
- For this particular job we do not wish to search for *NMR* or *DNA* structures so deactivate the check boxes adjacent to these options. And in this specific example, prior inspection of the site would have let you know that neither neighbouring protein chains nor additional ligands are involved so the *Contact filters* do not need to be activated.
- We do however wish to search only a kinase subset of the Relibase+ database and therefore need to select the tab marked **Hitlist Controls**.
- Select the *key_kinase* hitlist that you prepared earlier from the pull-down menu marked *Restrict search to hitlist named*.
- Enter an appropriate name for the search, e.g. `secstructtutorial`, in the *Save search in hitlist named* window and select *reli* from the list of databases provided in the *Search in database domains* window.
- In this particular search, you haven't specified any atoms onto which the search hits are to be superimposed and because this search is fairly fast (~15-25 mins) there is no need to run it as a batch job so the options below do not have to be switched on.

- Click on **Start** to start the search.
- When the search is finished, a new *Search Results* page will open.
- Use the *Save Search Results* option at the bottom left of this page to save the results of this search.
- The resulting hitlist is a list of Tyr/Ser/Thr kinase allosteric inhibitors. It will not be exhaustive though. A cursory inspection of the matches found might suggest a query that includes an essential lipophilic contact with the Cbatom of the DFG Asp and a good H-bond NH donor contact with an a-helical (SHAFT classified) glutamic acid residue (found on the mobile Ca-helix) and/or an H-bond acceptor contact with the same DFG Asp might produce a more extensive set of matches. Try this search out if you wish.



- If you are persuaded that filling the hydrophobic pocket newly created by the conformational change of the activation loop is essential for promoting this movement, look at the structure of 2p2i!
- We have a suggestion that the DFG out conformation is frequently associated with a type o.4 (VII3) open turn. Two other open turn types were also associated with the Phe and Gly residues of DFG in 1fpu (o.4 (IX) and o.4 (XII)). It might be worth redoing the constrained

substructure search again, but this time make sure the *Phe:A382* is constrained to be present in all three turn types. Do you get the same hits?

- What secondary structure features appear in ABL kinase inhibitor/protein structure which have DFG *in*, not *out*? You can investigate this by carrying out a **Similar Binding Sites** search, starting from the 1fpu *Ligand Information* page. Looking for high homology inhibited structures will allow you to identify from the overlay of these structures, other ABL kinase models with DFG *in*.
- If you would rather not do this search, have look at the secondary structures around the DFG region, of protein/ligand complexes 3dk3, 3dk6, 2hzi and f4j; and apo structure 2g2i. You should be able to identify for the inhibited structures one particular open turn which is not o.4 (VII3), and one or other of two closed turns, all of which are associated with the DFG loop and adjacent Leu:A384; and another very short and distinctive secondary structure element normally associated with Ser:A385 and Arg:A386.

This ends the tutorial